UCLA
UCLA Electronic Theses and Dissertations

Title

Beyond Taking Turns in Human-AI Interaction

Permalink

https://escholarship.org/uc/item/28d8b7wd

ISBN

9798273398702

Author

Liu, Bruce

Publication Date

2026-01-23

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Beyond Taking Turns in Human-AI Interaction

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Electrical and Computer Engineering

by

Bruce Liu

2026

© Copyright by

Bruce Liu

2026

ABSTRACT OF THE DISSERTATION

Beyond Taking Turns in Human-AI Interaction

by

Bruce Liu
Doctor of Philosophy in Electrical and Computer Engineering
University of California, Los Angeles, 2026
Professor Xiang Chen, Chair

Most human–AI systems today are built around a default paradigm of turn-by-turn interaction: users act, systems respond, and collaboration unfolds through discrete exchanges. While effective for many tasks such as conversational agents, this model shapes interaction to be linear rather than multidimensional, reactive rather than proactive, and transient rather than persistent.

Drawing on theories from communication and linguistics, this thesis argues that turn-taking has become an implicit and largely unexamined design constraint in human–AI interaction. I propose moving beyond turn-taking by introducing interaction structures that support parallel communication, proactive participation, and persistent shared spaces.

This thesis presents three interactive systems that instantiate and evaluate these structures in real-world contexts: *Visual Captions* demonstrates parallel communication by augmenting verbal conversations with real-time, AI-generated visuals. *Inner Thoughts* demonstrates proactive participation in multi-party settings. *Doki* demonstrates persistent shared space with a text-native interface for generative video authoring. Through system design, implementation, and empirical evaluation, this thesis shows how moving beyond turn-taking

can enable richer and more flexible forms of human–AI interaction, and outlines a set of design principles for building the next generation of interactive AI systems.

The dissertation of Bruce Liu is approved.

Haijun Xia

Nanyun Peng

Eunice Jun

Xiang Chen, Committee Chair

University of California, Los Angeles

2026

To my mother, Guizhen Du

and my father, Zeli Liu

for their unconditional love and support

TABLE OF CONTENTS

1 Introduction **1**

1.1 Thesis Contributions 4

1.2 Thesis Outline 5

2 From Linear to Parallel: Augmenting Verbal Communication with On-the-fly Visuals **8**

2.1 Introduction 8

2.2 Related Work 11

2.2.1 Text-to-Visual Systems 11

2.2.2 Proactive and Continuous AI Agent 12

2.2.3 Augmented Communication 13

2.3 Formative Study 13

2.3.1 Procedure 14

2.3.2 Design Space for Augmenting Verbal Communication with Visuals . . 14

2.4 VC1.5K Dataset 18

2.5 Model 21

2.5.1 Visual Intent Prediction Model 22

2.5.2 Technical Evaluation 23

2.6 Visual Captions System 24

2.6.1 Visual Captions Interface 26

2.6.2 Visual Captions Settings 28

2.6.3 AI Proactivity in Visual Captions 28

2.7 Evaluation 30
2.7.1 Pilot Study 30
2.7.2 Materials 31
2.7.3 Participants 31
2.7.4 Procedure 32
2.7.5 Findings: Using Real-time Visuals in Conversations (RQ1) 32
2.7.6 Findings: Interacting with AI (RQ2) 37
2.8 Case Study: Usage of Visual Captions In-the-wild 41
2.9 Discussion and Future Work 43
2.10 Conclusion 45

3 From Reactive to Proactive: Conversational Agents with Inner Thoughts 47

3.1 Introduction 48
3.2 Related Work 51
3.2.1 Proactive AI and Conversational Agents 51
3.2.2 Turn-taking in Multi-Party Conversations 52
3.2.3 Language, Thought, and LLM Agents 53
3.3 Next-Speaker Prediction Is Insufficient to Enable Proactive Conversational AI 54
3.3.1 Hypotheses 55
3.3.2 Materials & Methods 55
3.3.3 Results 57
3.3.4 New Task: Speaker Name or “Anyone” 57
3.4 Retrospective Think-aloud Study 58
3.4.1 Participants 58

3.4.2 Procedure . . . 58
3.4.3 Findings . . . 59
3.5 Inner Thoughts Framework . . . 63
3.5.1 Trigger . . . 64
3.5.2 Retrieval . . . 65
3.5.3 Thought Formation . . . 66
3.5.4 Thought Evaluation . . . 68
3.5.5 Participation . . . 69
3.5.6 Demonstration of AI's Proactive Behavior Enabled by Inner Thoughts 71
3.6 Simulative Evaluation . . . 74
3.6.1 Apparatus: the Inner Thoughts Playground . . . 74
3.6.2 Conditions . . . 76
3.6.3 Agent Personas and Conversation Generation . . . 76
3.6.4 Hypothesis . . . 77
3.6.5 Human Evaluation of Simulated Conversations . . . 78
3.6.6 Findings . . . 79
3.7 User Evaluation . . . 81
3.7.1 Apparatus: *Swimmy* Slackbot . . . 81
3.7.2 Study Design . . . 83
3.7.3 Findings . . . 84
3.8 Discussion and Future Work . . . 87
3.8.1 Proactive Agents via Intrinsic Motivation vs. Extrinsic Cues . . . 87
3.8.2 Applying Inner Thoughts Beyond Casual Conversations . . . 87

3.8.3 Proactive Conversational Agents Beyond Text and Computational Efficiency . . . 88
3.8.4 Technical Limitations of the Inner Thoughts Framework. . . . 88
3.8.5 Automatic Evaluation for AI Proactivity . . . 89
3.8.6 Other Applications of Inner Thoughts . . . 90
3.8.7 Interacting with the Thoughts of LLMs . . . 90
3.9 Conclusion . . . 91

4 From Transient to Persistent: A Text-Native Interface for Generative Video Authoring . . . 94

4.1 Introduction . . . 94
4.2 Related Work . . . 97
4.2.1 Generative Models as Individual Video Generators . . . 97
4.2.2 Novel Video Authoring Interfaces . . . 98
4.2.3 Rich-Text Editing and Dynamic Documents . . . 99
4.3 How People Create Generative Videos Today . . . 100
4.3.1 Motivating Scenarios . . . 100
4.3.2 Design Principles . . . 103
4.4 Doki . . . 104
4.4.1 Example Workflows . . . 105
4.4.2 A Structured Text Representation for Generative Videos . . . 106
4.4.3 Doki System . . . 107
4.4.4 Doki's Shot Generation Pipeline . . . 118
4.4.5 Implementation . . . 120

4.5 Diary Study . . . 121
4.5.1 Participants . . . 121
4.5.2 Procedure . . . 122
4.6 Results . . . 125
4.6.1 Doki Usage . . . 125
4.6.2 RQ1. How do users perceive the benefits and limitations of Doki's text-native interface? . . . 126
4.6.3 RQ2. What workflows do people employ when using Doki? . . . 131
4.6.4 RQ3. How do users' prior experience with video editing and generative AI influence the ways they perceive and use Doki? . . . 134
4.7 Discussion and Future Work . . . 136
4.8 Conclusion . . . 139

5 Conclusion . . . 140
5.1 Review of Thesis Contributions . . . 140
5.2 Future Work . . . 141

References . . . 143

LIST OF FIGURES

1.1 The ELIZA interface (1966), an early example of chat-based, turn-taking human–computer interaction. 2

1.2 Comparison of turn-taking interaction and three alternative human–AI interaction structures explored in this thesis. Blue rectangles represent user contributions and green rectangles represent AI contributions over time. In traditional turn-taking (top left), interaction proceeds as a strictly sequential alternation between user and system. *Visual Captions* (top right) introduces parallel communication channels, allowing human and AI contributions to overlap and unfold concurrently. *Inner Thoughts* (bottom left) relaxes reactivity by enabling the AI to generate internal thoughts outside explicit turns. *Doki* (bottom right) moves beyond transient exchanges by supporting dense, persistent interaction within a shared workspace, where human and AI actions interleave continuously over time. 3

2.1 Design space for augmenting verbal communication with visuals. Visual Captions focuses on generating synchronous captions with a wide range of visual content, type, and sources. Our lab study allows users to express oneself in one-to-one remote conversations, while we deploy Visual Captions in the wild to explore one-to-many (when a single individual sends messages to a large audience, *e.g.*, giving a presentation) and many-to-many (when multiple people communicate with each other simultaneously, *e.g.*, group social event) scenarios. Visual Captions also provides three levels of AI proactivity to accommodate different interaction preferences. 15

2.2 From our formative study, participants reported a number of use cases of Visual Captions and how visuals would facilitate communication. We list the reported themes and the number of participants who mentioned the theme (#P). 16

2.3 Crowdsourcing interface on MTurk we designed to collect dataset about people's preferences in augmenting verbal communication with visuals. In the crowdsourcing task, participants were asked to suggest visuals to supplement the *last sentence* in the conversations. We also show the previous sentence under *Previous*. 19

2.4 Distribution of the crowdworkers' responses of the appropriate visual sources and types. 19

2.5 Model performance over 1500 steps. 21

2.6 Technical evaluation results of the visual prediction model rated by crowdworkers. 23

2.7 Examples visual intent predictions by our model. Compared with keyword-based approaches, our system could handle open-vocabulary conversations and contextually predict visual content, visual source, and visual type. 25

2.8 System workflow of Visual Captions. The workflow consists of three major steps: A. Transcription Parsing; B. Visual Intents Prediction; C. Visual Generation. . . 27

2.9 In Visual Captions's interface (default auto-suggest mode), the *scrolling view* (A) displays privately candidates of visual suggestions generated by our visual intent prediction model. Emoji suggestions are displayed on the bottom right corner (C). Users can click and display the visual to the *spotlight view* (B) to share it publicly. 29

2.10 Participants' Likert scale ratings to (1) Knowledge: The suggested visuals can help me understand some unfamiliar terms and concepts; (2) Clarification: The suggested visuals can help clarify ambiguity in the conversation; (3) Engagement: The suggested visuals can make the conversation more interesting and enjoyable; (4) Creativity: The suggested visuals provided topics to talk about or led the direction of the conversation. 33

2.11 Participants' Task Load Index and Likert scale ratings (from 1 - Strongly Disagree to 7 - Strongly Agree) to four conversations with different Visual Captions modes: No VC, Auto-Display, Auto-Suggest and On-Demand Suggest, with 95% confidence interval bar. In the Task Load Index questions we asked participants their feeling of mental demand, physical demand, frustration (how stressed and annoyed), temporal demand (how hurried and rushed), and effort. In Likert scale questions we asked participants about their experience interacting with the AI, and their feeling of having control, AI intrusiveness, difficulty to override errors, and AI's interruption to the conversation. We additionally annotated statistical tests (Mann-Whitney U) results between different conditions: ns (non-significant), * (p<.05), ** (p<.01), *** (p<.001). 38

3.1 People's intrinsic motivation to engage in conversations: Heuristics of what factors influence people's decisions to express or withhold their thoughts during conversations, derived from our think-aloud study. Each heuristic contains two example mid-level themes from our codebook. 59

3.2 The *Inner Thoughts* framework for AI proactive engagement in conversations. A conversational event triggers the retrieval of relevant memories from long-term memory and thought reservoir. New thoughts are then formed based on these activated memories, and added to the thought reservoir. These thoughts are evaluated for AI's intrinsic motivation (score = 4.1 in the figure) to express. AI participates by articulating a thought at a selected moment in the ongoing conversation. 63

3.3 Prompt structure for evaluating *intrinsic motivation* of a thought. The evaluator rates the AI's intrinsic motivation to engage using a 1-5 scale based on heuristics like relevance and coherence. A Chain-of-Thoughts (CoT) process evaluates both positive and negative factors, resulting in a weighted score. 67

3.4 Examples selected from simulation logs of AI turn-taking behaviors in the Inner Thoughts framework. The figure illustrates four behaviors: Participation by Motivation, where the AI joins the conversation by sharing relevant personal experience; Interruption, where the AI interjects with a strong contribution during an ongoing discussion; Retention, where the AI holds back a thought until it's contextually relevant; and Thought Evolution, where the AI adapts its responses as the conversation progresses. 71

3.5 The Inner Thoughts playground web app interface. Multiple AI and humans can be added to simulate a group conversation. Users can also view and edit each of the participants' long-term memory and thoughts. 75

3.6 Stacked bar plots of participants' ratings on the metrics used in our technical evaluation to measure the quality of AI simulated conversations. Each statement is rated on a Likert-scale from 1 – Strongly Disagree to 7 – Strongly Agree. . . . 78

3.7 An example conversation between two human participants and *Swimmy*, a Slackbot developed based on the Inner Thoughts framework, on Slack. 82

4.1 A comparison of interface paradigms. (a) "Bento Box" style interfaces distribute authoring across multiple, separate representations. (b) Doki's approach uses a text-native canonical representation where the document serves as the primary interface. 98

4.2 Two basic example workflows in Doki. Alice: (1) define assets and shots with slash commands → (2) write story and generate previews → (3) create video shots; Bob: (a) prompt the sidebar agent for a draft → (b) review the AI-generated draft → (c) refine with inline AI agent. 104

4.3 Creating a shot in Doki. A new shot is inserted inline with a slash command, and the description that follows serves as its prompt. The system first generates a preview image, which can then be turned into a video clip. Users can click to expand them for playback and additional controls. 108

4.4 Writing consecutive shots within a single paragraph. Later shots inherit context from earlier ones, enabling continuity across a sequence. We achieve strong consistency between shots without repetitive context description in the Doki document. 109

4.5 Creating definitions in Doki. Users open the command menu with a `/`, select a type, and provide a name and description. Optionally they can add a visual definition for even better consistency. 112

4.6 Adding variants to a definition and shot. The user selects Add Variants from the context menu by right-clicking on the image. Three variations will be added inline. The user can then select one variant by moving it to the first position. . 113

4.7 Authors can apply definitions globally (#all =) or scope them to specific sections via headings. In the example, a scoped definition (#Cartoon) transforms subsequent shots under the heading level into a consistent cartoon style. 114

4.8 Three types of inline agent actions. (a) "Enhance" increases the descriptiveness of the selected text. (b) "Create Definition" converts the selection into a reusable definition. (c) "Custom Request" applies an in-context edit based on user instruction. 116

4.9 Doki's shot generation pipeline. First, the user's raw prompt and the document context is passed to reference resolution module to create a structured prompt and gather relevant visual reference images from the definitions. Then, the prompt rewriter rewrites and polishes this structured prompt for image and video generation. Doki first produces a static preview image and then uses that image as a the first frame to generate the final video clip. 118

4.10 Example frames from videos created by participants in our diary study. The videos demonstrate the breadth of video types Doki supports. Each row shows a video through four representative keyframes. 124

LIST OF TABLES

3.1 Next speaker prediction accuracy for different GPT models, grouped by the turn-taking type of the current utterance: self-selection (*Self*), turn-allocation (*Alloc*) and overall. The random guessing baseline accuracy is 0.127. 56

3.2 Study conditions for technical evaluation and user study. We compare Inner Thoughts with the baseline approach of deciding when to participate by next speaker prediction, and then generate response based on AI's persona 93

3.3 Metrics used in our technical evaluation to measure the quality of AI simulated conversations. Each statement is rated on a Likert-scale from 1 – Strongly Disagree to 7 – Strongly Agree. 93

4.1 System Usability Scale (SUS) scores and subscales (Usability and Learnability) across participant roles, video creation frequency, and generative AI tool use. . . 126

ACKNOWLEDGMENTS

First and foremost, I would like to express my gratitude to my advisor, Xiang 'Anthony' Chen, for his guidance and support throughout my Ph.D. journey. I would also like to thank my committee members, Haijun Xia, Nanyun Peng and Eunice Jun, for their valuable feedback and suggestions. None of the projects in this thesis would have been possible without the contribution of my awesome collaborators: Amy Pavel, Patrick Carrington, Ruolin Wang, Alex Olwal, Peggy Chi, David Kim, Jiahao Li, Shitao Fang, Weiyan Shi, and Chien-Sheng Wu. I am grateful to Ding Li, Mira Dontcheva, Takeo Igarashi, Ruofei Du, Rajan Vaish and Brian A. Smith, who are my mentors during my internships. I would also like to thank my friends who have supported me throughout my studies, including but not limited to Peiling Jiang, Rima Cao, Zixiong Su, Mingrui Zhang, Xiaoyi Zhang, Toby Li, Hyunsung Cho, Mina Huh, Daehwa Kim, Vimal Mollyn, Yudai Tanaka, Riku Arakawa, Ruei-Che Chang, Zheng Zhang, Jiaju Ma, Erzhen Hu, Michael Liu, Liang He, Bryan Wang. Last but not least, I would like to thank my partner Ruby Chou, and my parents for their love and support.

VITA

2016–2020	B.S. (Statistics and Machine Learning, Human-Computer Interaction) with University Honors, Carnegie Mellon University, Pittsburgh, Pennsylvania.
2018	Machine Learning Intern, Dell EMC.
2020–2022	M.S. (Electrical and Computer Engineering), UCLA, Los Angeles, California. Received the Distinguished Master's Thesis Research Award.
2021	Research Intern, Snap Research.
2022	Student Researcher, Google.
2023	Visiting Researcher, Igarashi Lab, The University of Tokyo, Tokyo, Japan.
2023	Recipient of the Amazon Ph.D. Fellowship.
2024	Research Scientist Intern, Meta Reality Labs.
2025	Research Scientist Intern, Adobe Research.
2025	Recipient of the UCLA Dissertation Year Award.
2020–2026	Author and co-author of over 10 peer-reviewed publications in human-computer interaction (CHI, UIST, CSCW, ASSETS) and holder of 8 US patents/applications.

CHAPTER 1

Introduction

The recent advancements in Large Language Models (LLMs) and Generative Artificial Intelligence (AI) have fundamentally altered the landscape of Human–Computer Interaction (HCI). We have moved from systems that could merely label an image as "dog" or "cat" to agents capable of reasoning, generation, and complex problem-solving. Yet, despite this leap in model capability, the primary vessel for human–AI interaction has barely changed: the chat box (Figure 1.1). Whether building apps, generating videos, or planning a trip, users overwhelmingly engage with these models through a rigid, sequential paradigm: the user prompts, the system waits; the user submits, the system responds. This is the paradigm of *turn-taking*.

Turn-taking is one of the most fundamental structures of human communication. As described in Sacks et al.'s seminal work *A Simplest Systematics* [114], conversation is organized through distinct turns that regulate the flow of speech and prevent overlap. This structure was naturally adopted in early computing interfaces, from Command Line Interfaces (CLIs) to later Conversational User Interfaces (CUIs), where clear alternation between user input and system output was both intuitive and technically necessary. Over time, turn-taking became embedded not only in our social expectations of interaction, but also in the technical foundations of modern AI systems. Large Language Models (LLMs), trained using autoregressive objectives that predict and generate one token at a time, are inherently oriented toward sequential forms of interaction.

Turn-taking is effective for many scenarios. It provides clarity, regulates information

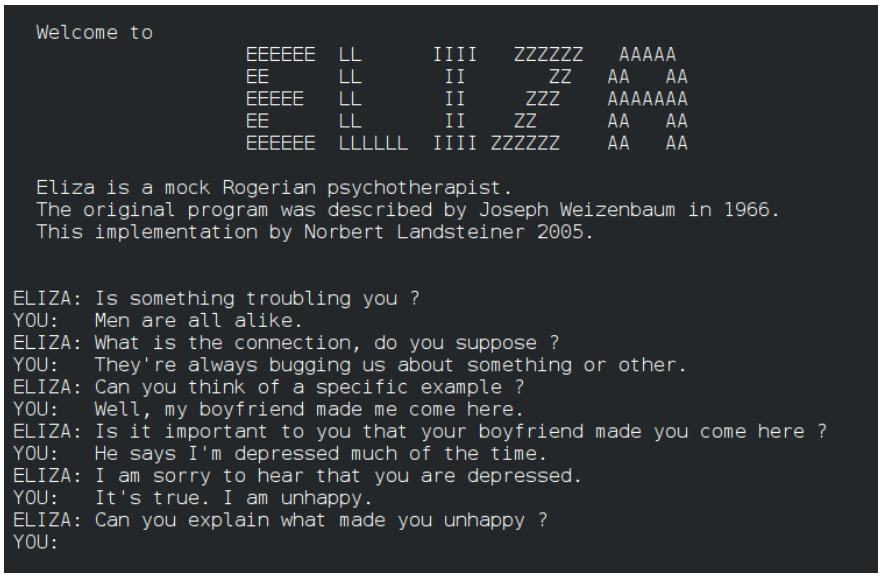


Figure 1.1: The ELIZA interface (1966), an early example of chat-based, turn-taking human–computer interaction.

flow, and aligns closely with familiar patterns of human dialogue. However, I argue that as this structure has become the default, it increasingly functions as an implicit and largely unexamined design constraint. Interactive AI systems are often designed as if the chat interface represents the upper bound of what interaction can be, rather than one point within a broader design space. As AI models advance, this assumption can limit new possibilities for how humans and AI systems might interact.

In their paper *Beyond Being There* [58], Hollan and Stornetta argue that computer-mediated communication systems should not merely replicate face-to-face interaction, but instead leverage the unique affordances of digital media to enable forms of communication that are physically *impossible* in the real world. This thesis advances an analogous argument in the context of human–AI interaction. By designing AI systems to mirror the turn-taking mechanics of human conversation, we inherit a set of constraints that may not be necessarily intrinsic to intelligent systems. Moving beyond turn-taking therefore opens a broader and potentially more exploratory design space. In this thesis, I examine three key constraints

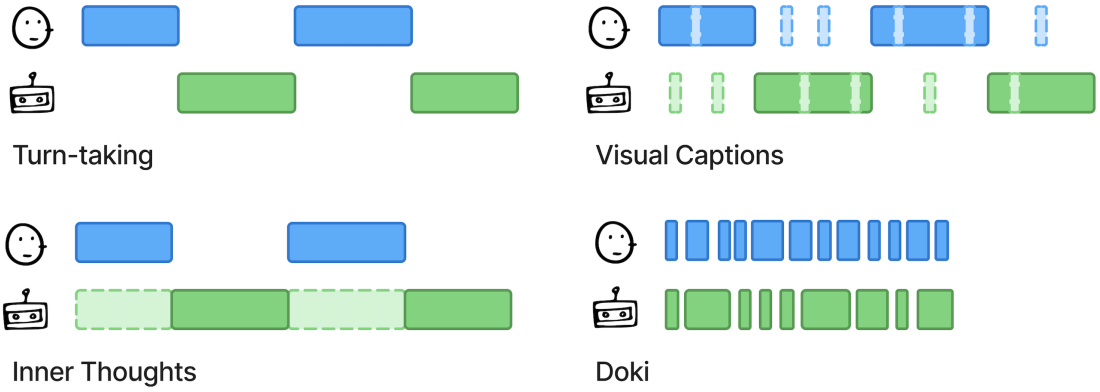


Figure 1.2: Comparison of turn-taking interaction and three alternative human–AI interaction structures explored in this thesis. Blue rectangles represent user contributions and green rectangles represent AI contributions over time. In traditional turn-taking (top left), interaction proceeds as a strictly sequential alternation between user and system. *Visual Captions* (top right) introduces parallel communication channels, allowing human and AI contributions to overlap and unfold concurrently. *Inner Thoughts* (bottom left) relaxes reactivity by enabling the AI to generate internal thoughts outside explicit turns. *Doki* (bottom right) moves beyond transient exchanges by supporting dense, persistent interaction within a shared workspace, where human and AI actions interleave continuously over time.

imposed by the default turn-taking paradigm on human–AI interaction:

- *Linearity*: Turn-taking structures interaction as a single-threaded sequence, constraining communication to one stream of action and response at a time.
- *Reactivity*: Turn-taking positions the AI primarily as a responder, with system behavior driven by explicit user input rather than proactive contribution.
- *Transience*: Turn-taking frames collaboration as a sequence of exchanges, where context is captured only in a chat log and quickly recedes into the past.

This thesis explores the design space beyond turn-taking. Drawing on theories from communication, linguistics, and Computer-Supported Cooperative Work (CSCW), it proposes

and evaluates a set of interaction structures that move from linear to parallel communication, from reactive to proactive participation, and from transient exchanges to persistent shared spaces. Through the design, implementation, and empirical evaluation of three human–AI systems, *Visual Captions*, *Inner Thoughts*, and *Doki*, this work shows how relaxing the turn-taking constraint allows humans and AI to communicate through multiple simultaneous channels, act without waiting for explicit prompts, and collaborate within shared spaces that accumulate and organize ongoing work rather than reducing interaction to a sequence of isolated exchanges. Together, these systems show how moving beyond turn-taking can enable richer and more flexible forms of human–AI interaction, and outline a set of design principles for building the next generation of interactive AI systems.

1.1 Thesis Contributions

This thesis makes three main contributions to the field of HCI and human–AI interaction.

First, it introduces a conceptual framing of *turn-taking* as an implicit and largely unexamined design constraint in contemporary human–AI systems. While turn-taking has been widely adopted due to its roots in human conversation and early computing systems, this work articulates how its assumptions (*e.g.*linearity, reactivity, and transience) shape and limit interaction. This thesis provides a lens for critically examining existing human-AI interfaces and for imagining alternative interaction paradigms.

Second, the thesis proposes a set of interaction structures that move beyond turn-taking, grounded in theories from communication, linguistics, and CSCW. These structures explore parallel communication, proactive engagement, and persistent shared spaces as alternatives to single-threaded, reactive chat interactions.

Third, the thesis provides empirical evidence for the value of these interaction structures through the design, implementation, and evaluation of three human–AI systems: *Visual Captions*, *Inner Thoughts*, and *Doki*. Through qualitative and quantitative studies, these

systems demonstrate how relaxing the turn-taking constraint can enable richer forms of collaboration, support ongoing sensemaking, and reshape users' mental models of AI from passive respondent to active participant. These findings inform a set of design implications for building future interactive AI systems that extend beyond the chat-based paradigm.

1.2 Thesis Outline

This dissertation proceeds as follows.

Chapter 2: Visual Captions (Augmenting conversations with parallel visuals). Chapter 2 introduces *Visual Captions*, a system that enriches live conversation by suggesting and sharing relevant visuals in video conferences. Visual Captions builds on a crowdsourced dataset (VC1.5K) that contains 1595 language–visual annotations and supports open-vocabulary visual suggestions during speech. The chapter describes the end-to-end workflow (transcription parsing, visual intent prediction, and visual generation), and the interaction design that supports different levels of AI proactivity (auto-display, auto-suggest, and on-demand-suggest). Finally, Chapter 2 reports user findings on how participants experienced and preferred different proactivity levels (with substantial variance across users), and how proactivity trades off control, effort, and interruption in synchronous communication.

Chapter 2 is a version of the published paper, Visual Captions: Augmenting Verbal Communication with On-the-fly Visuals. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 108, 1–20. https://doi.org/10.1145/3544548.3581566

Chapter 3: Inner Thoughts (Proactivity via a parallel thought stream). Chapter 3 presents *Inner Thoughts*, a framework for proactive conversational agents in multi-party conversation. The core idea is to run a continuous, covert train of thoughts in parallel to the

visible conversation and decide when to speak by scoring the agent's intrinsic motivation to express a thought. The chapter first grounds this approach in a formative think-aloud study that derives engagement heuristics (*e.g.*, relevance and information gaps) and operationalizes them into intrinsic motivation levels. It then details the five-stage framework (Trigger, Retrieval, Thought Formation, Evaluation, Participation), followed by system implementations and evaluations showing that Inner Thoughts improves perceived turn appropriateness and coherence relative to a next-speaker prediction baseline.

Chapter 3 is a version of the published paper, Proactive Conversational Agents with Inner Thoughts. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 184, 1–19. https://doi.org/10.1145/3706598.3713760

Chapter 4: Doki (A persistent document as the shared working space). Chapter 4 introduces *Doki*, a text-native interface for generative video authoring that treats a document as the primary substrate for ideation, structure, and production. Doki uses a hierarchical representation that aligns writing structure with video structure (document as video, paragraph as sequence, sentence as shot), and implements a minimal interface with inline shot elements and a slash menu. The chapter describes how Doki supports consistency and reuse through definitions and context propagation, and reports a week-long diary study with 10 participants that characterizes real workflows and identifies benefits such as rapid idea-to-content iteration and improved comprehension of narrative structure.

Chapter 4 is a version of the manuscript, A Text-Native Interface for Generative Video Authoring, in preparation for submission to an academic conference.

Chapter 5: Conclusion. Chapter 5 synthesizes findings across the three systems to articulate design implications for moving beyond turn-taking. It summarizes how parallel channels (Visual Captions), proactive internal processes (Inner Thoughts), and persistent

shared representations (Doki) relax the constraints of linearity, reactivity, and transience, and it outlines limitations and future research directions.

CHAPTER 2

From Linear to Parallel: Augmenting Verbal Communication with On-the-fly Visuals

This thesis argued that turn-taking imposes three often-unexamined constraints on human–AI interaction: it enforces a single linear channel of exchange, positions the AI primarily as reactive, and leaves collaboration transient and hard to act upon beyond the chat log. In Chapter 2, the first project of this thesis begins by relaxing the *linearity* constraint. Instead of concentrating all meaning in a single stream of spoken turns, we explore what happens when AI adds an additional, concurrent channel that can carry supportive content without requiring the speaker to stop, rephrase, or explicitly request help.

Visual Captions operationalizes this idea in synchronous conversation by providing on-the-fly visuals (e.g., images and emojis) that run in parallel with speech and can be surfaced with varying degrees of AI initiative. In deployment, participants described using the parallel visual channel to express additional meanings *without interrupting* whoever currently holds the floor. These motivations set up the chapter's central question: how can AI augment live conversation through parallel communication, and what does that reveal about interaction beyond a strictly turn-based chat paradigm?

2.1 Introduction

Recent computer-mediated systems are increasingly facilitating verbal communication, which is often the predominant mode of communication [144, 6]. Platforms like Google Meet,

Zoom, and Microsoft Teams, have been widely adopted and provided capabilities such as live captioning and noise cancellation to facilitate conversations.

While such tools help people understand each other better, we envision that visual augmentations that leverage the semantics of spoken language could also be helpful, especially for conveying complex, nuanced, and unfamiliar information. People are already using visual aids to provide additional context and clarification in daily conversations. For example, when talking about a recent trip, people may use photos from their album to help their listeners follow along. Similarly, when describing a new restaurant to a friend, one might say "*it's a small place with a lot of character*" and then search for a picture online to show what the place looks like. Research has shown that people learn more effectively from videos than from audio-only versions of the same material [88, 128, 67], and prefer podcasts [147] and stories [76] with visuals over those without. The Multi-modal Phenomena [90, 123, 127] and the Principle of Inverse Effectiveness [124] have also proved that the human sensory system has a superadditive effect when responding to stimulus from multiple, simultaneous modalities. As the adage goes, *"a picture is worth a thousand words"*.

Prior research has proposed various automated text-to-visual systems, that can automatically transform audio-only content into audiovisual content [76, 147], and create realistic images and art from natural language [156, 92]. However, augmenting *synchronous human-human verbal communication* with visuals presents unique challenges that have not been addressed by these systems.

First, the input is a continuous stream of conversation, rather than a discrete textual description of the visual, requiring the system to go beyond keyword-based approaches, like named entity detection, to understand the *implicit* intent of what people want to show in the context. *Second*, when people are actively engaged in conversation, they have limited cognitive resources to interact with AI prompts and results, making it necessary for the interaction to be subtle and minimal to avoid disrupting the conversation. *Third*, without a real-time system deployed, it is difficult to study how people could interact with and benefit

from visuals in real conversations, and how such systems would impact their communication.

In this work, we introduce Visual Captions, a system designed to address the challenges of augmenting synchronous human-human verbal communication with visuals. Visual Captions automatically predicts the "visual intent" of a conversation, or or the visuals that people would like to show at the moment of their conversation, and suggests them for users to immediately select and display. We conducted two rounds of formative studies with 10 participants and collected, annotated, and analyzed over 1,500 dialogues from 246 crowdworkers to understand the interest in such capabilities. Using this dataset, we trained an accurate, robust, and open-vocabulary language model to predict "visual intents" in conversations, achieving an 86.59% validation token accuracy. Based on the language model, we implemented Visual Captions as a user-customizable Chrome plugin with three levels of AI proactivity: *Auto-display* (AI autonomously adds visuals), *Auto-suggest* (AI proactively recommends visuals), and *On-demand-suggest* (AI suggests visuals when prompted).

We evaluated Visual Captions with 20 participants in a controlled lab study where participants chatted casually in four different scenarios, Participants found real-time visuals facilitated live conversations in multiple ways, including helping to explain and understand unfamiliar concepts, clarify language ambiguities, and make conversations more engaging. Participants also reported diverging preferences in how to interact with the AI in-situ. We also conducted a two-week in-the-wild deployment study with 10 participants to allow people to use the system in their everyday conversations. Participants reported that different levels of AI proactivity in Visual Captions were preferred in various social scenarios.

In summary, we contribute:

- A dataset of 1595 visual intents collected from 246 crowd workers, covering 15 categories of topics.
- A design space for real-time visual augmentation of verbal communication, derived from findings from two brainstorming sessions in a formative study (N=10).

- A visual intent model that predicts the type, source, and content for the visuals that people may want to display in conversations.
- Visual Captions, an open-source Chrome plugin for real-time augmented visuals in video conferencing products[1].
- Results from a user study (N=26) and a deployment study (N=10) that demonstrates the potential benefits of using visuals to augment conversations.

2.2 Related Work

Our work is inspired by prior research in automatic text-to-visual enhancement, proactive AI agents, and augmented communication.

2.2.1 Text-to-Visual Systems

Researchers have developed various computational techniques to use visual materials to support text-based contents [104, 73, 105, 115, 153, 158]. Early work enhanced emails by suggesting relevant personal photos related to the email topics [71]. As text chats became more common, real-time suggestions were developed to assist users in adding illustrations [66], images [32], and animated graphics [157, 146, 80]. For a comprehensive review of text-to-visual systems, we refer readers to the surveys by Hassani et al. [54] and Zakraoui et al. [154].

Recent research in HCI converts text articles to audio-visual slideshows [76, 28] and automatically generates visuals for travel podcasts [147] that consider sequential presentation and storytelling. When the source documents contain multimedia materials, automatic methods could crop, zoom, or loop images and videos to enhance text content [31, 30]. These methods demonstrated novel approaches for enriching text-driven experiences. Our work builds on

[1]Visual Captions is open-sourced at: `https://liubruce.me/visual_captions`

these systems but for face-to-face conversations, and providing real-time visuals relevant to the spoken content.

2.2.2 Proactive and Continuous AI Agent

Prior literature has investigated both proactive and continuous AI agents. In 1996, Rhodes et al. [108] propose one of the earliest system for continuously providing relevant information by monitoring human activities. However, AI is never perfect in real-world deployment [71]. Meurisch et al. [91] present an in-the-wild study of proactivity levels in voice-controlled assistants. Andolina et al. [5] present SearchBot, which provides continuous recommendations of related documents and entities in a non-intrusive way [4] during voice conversations. Their work leveraged Google's Cloud Natural Language API for extracting recognized entities from the transcripts. In contrast to their work, our system fine-tuned a large language models and better understand conversations with visual intent. We elaborate the differences between our approach and key-word based approach in Section 2.5.

In addition, mixed-initiative interactions, where both the user and the computer can initiate actions and take turns controlling the flow of the interaction, have been a topic of interest in HCI research. Horvitz et al.[60] reviewed critical factors for the effective integration of human and AI. Our work is inspired by prior research in mixed-initiative interaction and proactive AI. We provide three levels of AI proactivity and investigate the effects of these levels on users' perception and acceptance of the AI agent in a real-world setting. In addition, we study people's preferences on mixed-initiative interactions when they are actively engaged in conversations that limit their cognitive load and interaction bandwidth.

2.2.3 Augmented Communication

Our work is related to systems that augment verbal communication with text, interactive graphics, and computer-assisted actions. With recent advances of real-time machine learning techniques and wearable displays, there is a grouping trend of integrating augmented communication in everyday conversations [56]. Lyons et al. [87] propose Dual–purpose Speech, which automatically use speech to navigate a user's calendar, save transient information, and create asynchronous tasks. Saquib et al. [117] enable interactive presentations with body-driven graphics, by mapping a variety of body movements to a wide range of graphical manipulations beforehand. Müller et al. [94] present Cloudbits, which interactively renders calendar, emails, hotel, and restaurant during conversations via a Wizard-of-Oz study on an HMD. Peng et al. [100] present SpeechBubbles, a real-time speech-to-text system that associates utterances with speakers. Their studies show that participants significantly prefer their layout over traditional captions. Ogawa et al. [96] propose a smartwatch-based approach to suggesting topics in casual conversations. Social messaging apps like Snapchat and Facebook Messenger also suggest pre-defined stickers and emojis through keyword detection.

Visual Captions instead uses a fine-tuned large language model to proactively suggest relevant visuals in open-vocabulary conversations. Our system provide personalized visuals in real-time that are tailored to the specific context. To the best of our knowledge, Visual Captions is the first real-time system trying to understand user's visual intent, suggest relevant visual content, source, type, offers different levels of proactivity, and gets deployed in the wild without the limitation of pre-scripted conversations.

2.3 Formative Study

To understand how people would like to augment their speech with on-the-fly visuals, we conducted two brainstorming sessions. These results informed a design space for visually augmenting conversations.

2.3.1 Procedure

We recruited 10 participants via group email invitations and internal communication channels in Google. Participants had various technical and non-technical backgrounds, including software engineers, researchers, UX designers, visual artists, students, etc. We held one-hour brainstorming sessions with two groups of five participants. In each session, we introduced the low-fidelity prototypes of the envisioned system, followed by video demos of the existing text-to-image systems. Participants then brainstormed ideas on a digital white board based on three prompts: (1) What are some scenarios where you could imagine using real-time visuals to augment conversations? (2) What types of visuals would you like to add in conversations? (3) How would you like to present and interact with the visuals?

2.3.2 Design Space for Augmenting Verbal Communication with Visuals

Two researchers organized participants' responses with the affinity diagram approach. Informed by the set of low-level and high-level themes derived, we developed a design space for systems that augment verbal communication with visuals. We followed the design space analysis methods [24] and held iterative discussion sessions. We identified eight key dimensions as detailed in Figure 2.1.

D1. Temporal. To augment verbal communication with visuals, systems can be either *synchronous* or *asynchronous*. The majority of prior systems provides augmentations asynchronously. Users either have to set up corresponding visuals before (*e.g.*, pre-configure visuals for an upcoming presentation, and trigger visuals by gestures and keywords [117]), or select and edit visuals after the text is composed [76, 147]. Our system falls in the paradigm of synchronous augmentation, where users select appropriate visuals on-the-fly while engaging in conversations.

D2. Subject. Visual augmentations of spoken language can either be used by the speaker to express their ideas (visualize their own speech) or by the listener to understand others

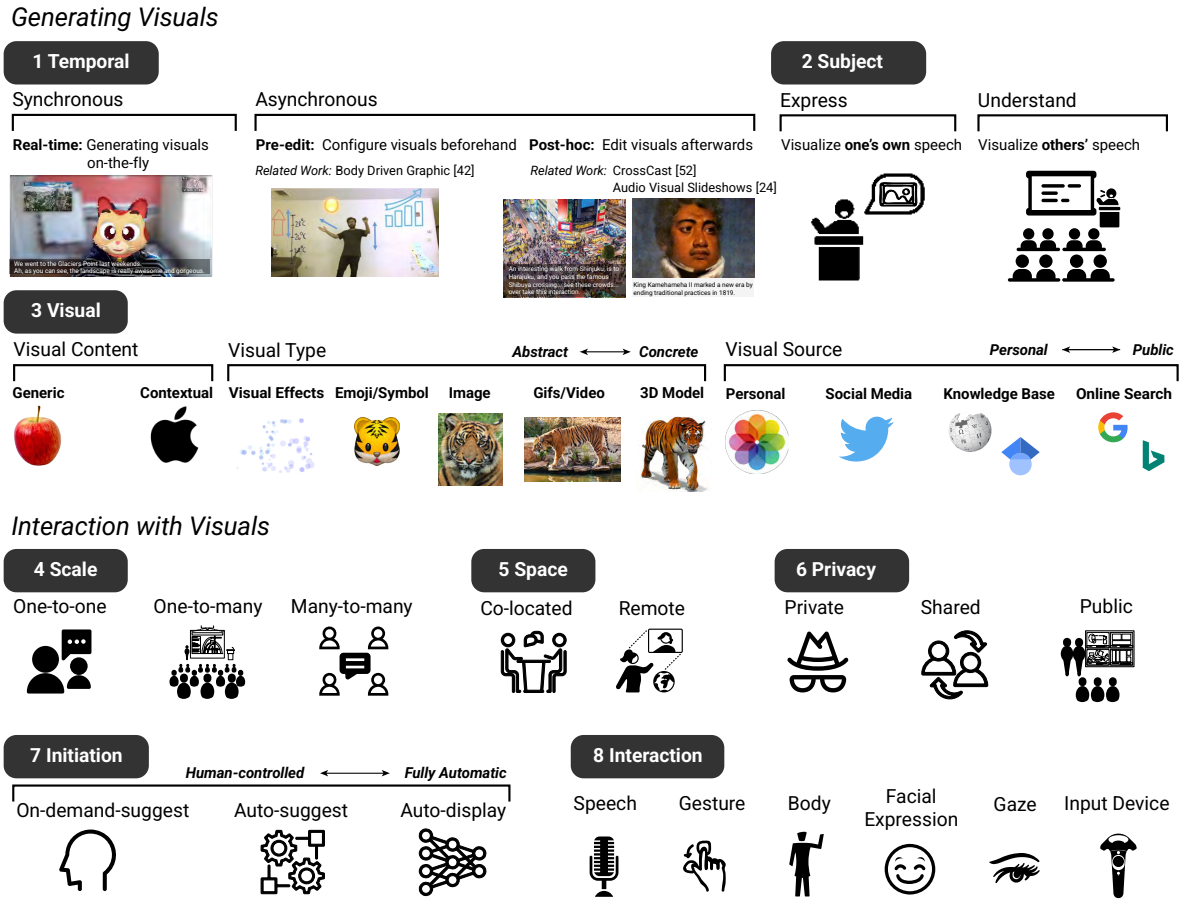


Figure 2.1: Design space for augmenting verbal communication with visuals. Visual Captions focuses on generating synchronous captions with a wide range of visual content, type, and sources. Our lab study allows users to express oneself in one-to-one remote conversations, while we deploy Visual Captions in the wild to explore one-to-many (when a single individual sends messages to a large audience, *e.g.*, giving a presentation) and many-to-many (when multiple people communicate with each other simultaneously, *e.g.*, group social event) scenarios. Visual Captions also provides three levels of AI proactivity to accommodate different interaction preferences.

(visualize others' speech). The majority of prior art in this domain falls under the former paradigm, where speakers select and design corresponding visuals to support their speech. In our system, we aim to support both subjects and allow all parties to visually supplement their own speech and ideas.

Use Cases & Scenarios	#P
Education and Lectures	10
• Math&Sciences: *"A math teacher demonstrates an orientational relationship between objects."*	
• History: *"History teachers could present Napolean when it comes to famous battles in Waterloo."*	
• Language Learning: *"Visualize unfamiliar words when learning a new language"*	
Casual Chats	10
• Introducing Pets: *"When talking about my dog show a picture for introduction."*	
• Unknown Dishes: *"When Ordering food at a Japanese Restaurant, see what a Sukiyaki is."*	
• Movie Information: *"Show movie posters when talking about recent movies."*	
Business and Utility	8
• Visual Navigation: *"Show directions with images of buildings."*	
• People's Names: *"Visualize people's profile in meetings."*	
• Mind Map Search: *"Have idea of something in mind but forgot - work with AI to find right imagery."*	
Creativity	5
• Project Brainstorming: *"Provide immediate visual materials for brainstorming to spur creativity."*	
• New Topics: *"Expand upon visuals to find new topics."*	
Storytelling	3
• 3D Animals: *"When telling a story to children, display 3D models of animals"*	

How Real-time Visuals Help		#P
Information	Provide and receive more information and new knowledge	10
Clarification	Resolve ambiguity and misunderstandings	10
Search	Quickly retrieve visuals from online search or photo album	8
Engagement	Make verbal communication richer and more interesting	6
Accessibility	Easier to understand for people with disabilities / with language barriers	5
Inspiration	Provide new topics and move the conversation forward	5
Memorization	Better memoriza content based on visuals	2

Figure 2.2: From our formative study, participants reported a number of use cases of Visual Captions and how visuals would facilitate communication. We list the reported themes and the number of participants who mentioned the theme (#P).

D3. Visual. Participants in our formative study wish to augment speech using a variety of visuals. We identified three main aspects to consider when providing visual augmentations:

(1) Visual Content — what information to be visualized? A segment of speech contains different information that can be visualized. For example, consider the statement *"I went to Disneyland with my family last weekend"*. One could visualize the generic term *Disneyland*, a picture of *I*, or more specific, contextual information such as *me and my family at Disneyland*. The system should be able to disambiguate the most critical and relevant information to visualize in the current context.

(2) Visual Type — how should the visual be presented? There are often multiple ways to present a visual, ranging from abstract to concrete. For example, the term *Disneyland* could be visualized as: an icon of Disneyland, a photo of Disneyland, an interactive 3D map

of Disneyland, or a video of people riding a roller-coaster. While visuals may have similar meaning, they can evoke different levels of attention and provide different levels of detail.

(3) Visual Source — where the visual should be retrieved from? Different sources can be utilized for the visual augmentations, including both personal and public assets. One might want to retrieve personal photos from one's own phone, or public images from the internet. While personal photos provide more contextual and specific information, images from the internet can provide more generic information with less privacy concerns.

Our system leverages a large language model optimized to consider the context of conversations, and identify the most appropriate visual content, type and source to suggest.

D4. Scale & D5. Space. Visual augmentations could be used in various communication scenarios, including one-on-one meetings, one-to-many lectures and many-to-many discussions. The number of participants and their location (*e.g.*in-person vs. remote) can greatly affect best practices for such visual augmentations. We developed Visual Captions for existing video conferencing software to augment meetings at different scales, supporting one-on-one, one-to-many, many-to-many scenarios.

D6. Privacy. Visual augmentations should take privacy into consideration right at the beginning and allow users to select among multiple privacy options: 1) Privately shown visuals are only presented to the speaker, invisible to any audience. 2) Publicly shown visuals are presented to everyone in the conversation. 3) In-between, the visuals can be selectively presented to a subset of audiences. We provide speakers with options 1) and 2) and to privately preview the visualizations before displaying them to the audiences. Listeners could also use our system to privately see the visuals based on speech they hear.

D7. Initiation. 6 participants in our formative study wanted the least efforts to generate visuals during the conversation, and therefore prefer having the system proactively providing visual augmentations without user interaction. However, other participants would like to have more control over the visuals, including when to trigger them and what to show. To

meet these different preferences, we designed three levels of AI proactivity: on-demand-suggest, auto-suggest, and auto-display.

D8. Interaction Participants mentioned six domains of potential interactions: speech (*e.g.*, *"let's show an image here"*), gesture (*e.g.*, pinch), body pose (*e.g.*, waving hands), facial expression (*e.g.*, to trigger emojis), gaze (*e.g.*, for selecting visuals from suggestions), and custom input devices (*e.g.*, a controller). We support traditional input devices (*e.g.*, keyboard, mouse, and touch screens) given their universal use in video conferencing. In Visual Captions, we trigger visual generation by understanding the language via speech-to-text engines or by user pressing the space bar. In the future, we can expand capabilities to enable interaction with body pose, facial expressions, and other devices.

During the session, participants discussed various potential use cases for Visual Captions and how they believe it would be helpful (Figure 2.2). Many participants expressed a desire to use Visual Captions for educational and casual scenarios, and said that they would appreciate the addition of visuals to make conversations more informative and clear.

2.4 VC1.5K Dataset

To further understand how people would prefer to augment their verbal communication with visuals, or their *visual intents*, we collected a dataset of 1595 sentence-visual pairs across a wide range of contexts. The VC1.5K dataset is available at: `https://liubruce.me/visual_captions`.

We collected a total of 1922 sentences (transcribed speech) from 42 YouTube videos (1201 sentences) and the DailyDialog dataset [79] (721 sentences). We manually categorized each video source and added a brief description of its context (*e.g.*, *best electronic products in 2021*). Videos were selected from various topics ranging from tour guide, lectures to documentaries etc. For each video, we downloaded its automatic captions and segmented

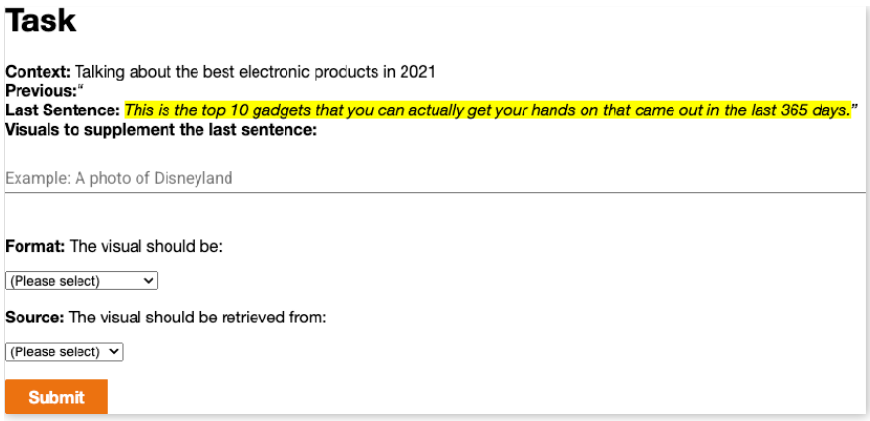


Figure 2.3: Crowdsourcing interface on MTurk we designed to collect dataset about people's preferences in augmenting verbal communication with visuals. In the crowdsourcing task, participants were asked to suggest visuals to supplement the *last sentence* in the conversations. We also show the previous sentence under *Previous*.

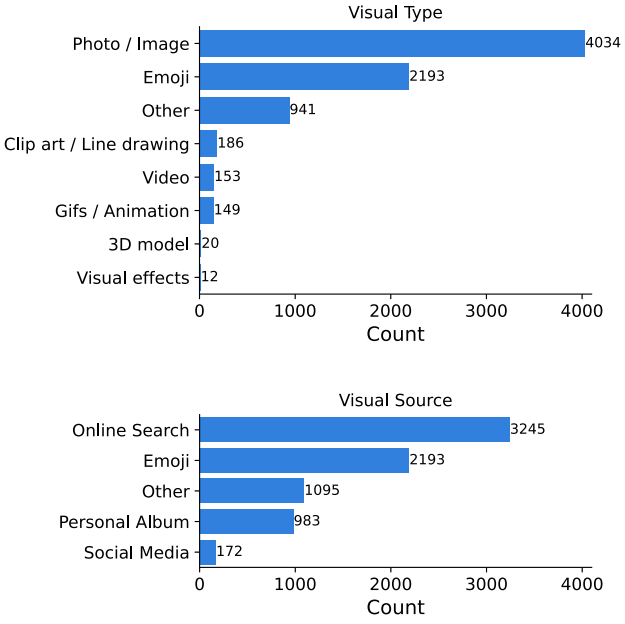


Figure 2.4: Distribution of the crowdworkers' responses of the appropriate visual sources and types.

the transcriptions into sentences using NLTK's Sentence Tokenizer[2]. Our dataset has few errors in sentence segmentation because all captions contain punctuation. We observed some short sentences (*e.g."Pop!"*), and mitigate this by also providing the previous sentence in both the labeling and training process.

We then crowdsourced visual intents through Amazon Mechanical Turk (MTurk). We provided specific task instruction and examples to crowdworkers before they start the task. For each annotation, we provided a short description of the video context, as well as the previous two sentences of the conversation. The crowdworkers were asked to *"Determine what visual content could be shown to supplement the last sentence, given context and previous conversation"*, and write their answer in the format of *Visual Type* of *Information To Visualize*. We additionally asked crowdworkers to select an appropriate visual type and visual source from a list of categories we identified from our formative study (Figure 2.3). Crowdworkers were allowed to submit multiple visuals per sentence. If there was no relevant visual content, they were asked to submit "none". Each sentence was annotated by 4 different crowdworkers, and workers were allowed to work on multiple HITs. We selected workers to be within the United States and have a history approval rate beyond 95%. We paid crowdworkers $0.15 for each HIT completed. A total of 246 unique crowdworkers were employed.

We post-processed the raw data to ensure that the annotations were in the correct format and that the crowdworker responses were consistent. Specifically, we checked that the visual type mentioned in the answers matched the visual type selected by the worker. We rejected and republished any tasks that did not pass these quality checks.

As our goal is to obtain a ground truth for what people want to show visually, we combined crowdworker responses by sentence similarity. Specifically, we first removed all the English stop-words and punctuations, and used NLTK's WordNetLemmatizer[3] to lemmatize

[2]NLTK's Sentence Tokenizer: `https://www.nltk.org/api/nltk.tokenize.html`

[3]https://www.nltk.org/_modules/nltk/stem/wordnet.html

the remaining words and transformed all the words to lowercase. We then used Sentence-Transformers [106] to compute two sentences' similarity. We consider two visual intents to be the same if their similarity score is greater than 0.85. After merging, 1595 sentences in our dataset have visual intents agreed by at least two crowdworkers, and 682 sentences by at least three crowdworkers. We removed all the sentences with less than two agreed visual intents. The final dataset (VC1.5K) consists of 1595 examples. We visualize the distribution of visual sources (Figure 2.4), and visual types (Figure 2.4) crowdworkers annotated.

2.5 Model

We trained a visual intent prediction model that predicts what visuals would be appropriate to show to supplement the ongoing conversation. After experimenting with different methods, including named entity detection, emotion analysis, and knowledge graph search, we used GPT-3 [22] fine-tuned with our VC1.5K dataset. This allows for a contextual system that can work on a range of topics without needing to be trained on specific vocabulary. We describe our training process, empirical analysis of model capabilities, and a technical evaluation.

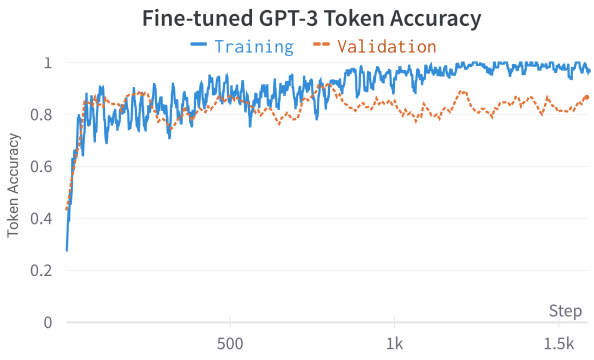


Figure 2.5: Model performance over 1500 steps.

2.5.1 Visual Intent Prediction Model

For training, we parsed each response into the format of "`<Visual Type>` of `<Visual Content>` from `<Visual Source>`". This serves as the label of our model. If there are more than one eligible label, we concatenate all responses together with a separator "`;`". If crowdworkers did not annotate any visuals for the sentence, we label "none". The prompt for each example is set to the previous two sentences in the text data. We additionally added a fixed separator, "$\rightarrow$", to inform the model when the prompt ends and the completion begins; and a fixed stop sequence, "`\n`", to indicate when the completion for the current example ends. In summary, we template the training data as:

{"prompt": "`<Previous Two Sentences>` $\rightarrow$",
"completion":
"`<Visual Type 1>` of `<Visual Content 1>` from
`<Visual Source 1>`;
`<Visual Type 2>` of `<Visual Content 2>` from
`<Visual Source 2>`;
...
$\backslash n$"}

We used 1276 (80%) examples for training and validation, the remaining 319 (20%) examples as test data. Data in each of the training, validation and test sets were selected from different videos to prevent overlaps. We fine-tuned the `text-davinci-002` model[4] for 4 epochs with a batch size of 8, at a learning rate of 0.05. API access to our model is available at `https://liubruce.me/visual_captions`.

We measured the performance of the fine-tuned model with the *token accuracy* metric, *i.e.*, the percentage of tokens in a batch that were correctly predicted by the model. In the

[4]OpenAI's Fine-tuning API: `https://beta.openai.com/docs/guides/fine-tuning`

process of training we did not perform any hyper-parameter optimization. During training, our model reached a training token accuracy of 96.81%, and a validation token accuracy of 86.59% (Figure 2.5). We additionally evaluated the categorical prediction accuracy of visual type and visual source on the test set. Our model reached an 87.6% accuracy for visual type and an 86.1% accuracy for visual source predictions.

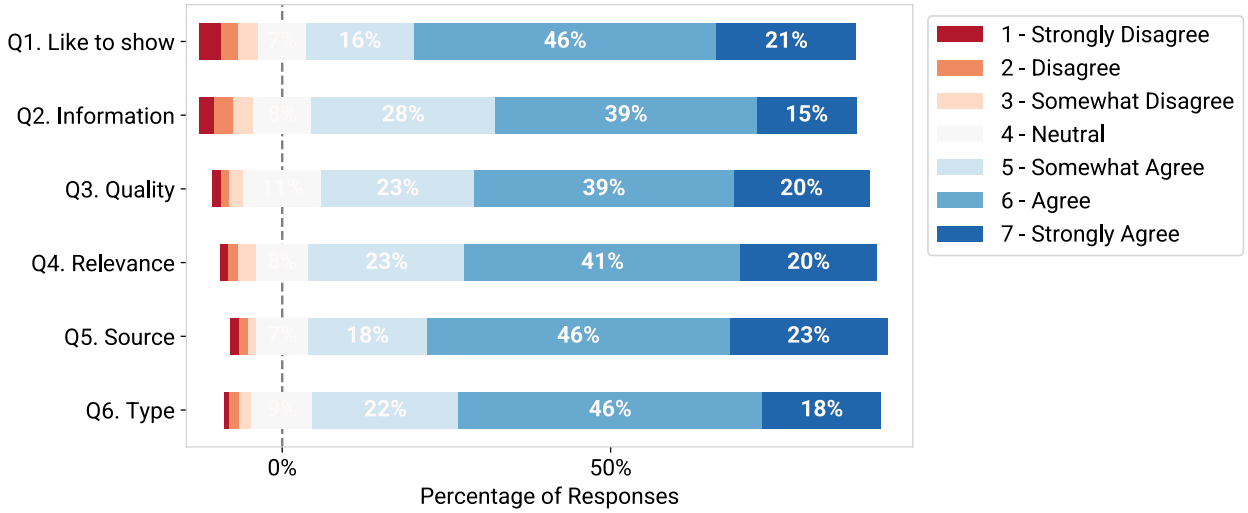


Figure 2.6: Technical evaluation results of the visual prediction model rated by crowdworkers.

2.5.2 Technical Evaluation

To better understand the generality of our model and the quality of visual suggestions, we conducted a technical evaluation with crowdworkers recruited on MTurk. We first generated visuals for the out-of-bag test dataset using our model. 282 visuals were generated from the 266 (out of 319 sentences in total) examples that had at least one visual suggestion (*i.e.*, not "none"). In this evaluation, we retrieved all visuals online, and provided crowdworkers with the original sentence, predicted visual content, visual type, and visual source. We asked crowdworkers to rate the visual suggestions on four dimensions using a 7-point Likert-scale from *1–Strongly Disagree* to *7–Strongly Agree*. Each example was evaluated by three crowdworkers (a total of 846 tasks). As an attention test, we also asked crowdworkers

to briefly explain their reason for the first question. We examined crowdworkers' responses, rejected and republished 139 low quality results that were too short, ambiguous, or irrelevant. We selected workers with approval rates over 95% and a US-based location, and paid them $0.30 per task completed. A total of 89 workers were employed.

Overall, our model and visual suggestion pipeline produced desirable visuals (Figure 2.6). Crowdworkers rated the statement *"I would like to display the visual when having the conversation (Q1)"* with a greater than or equal to *5–Somewhat Agree* 83% of the time. They consider the displayed visuals to be useful and informative (Q2, 82% $\geq$ *5–Somewhat Agree*), high-quality (Q3, 82% $\geq$ *5–Somewhat Agree*), and relevant to the original speech (Q4, 84% $\geq$ *5–Somewhat Agree*). In addition, crowdworkers found the predicted visual type (Q5, 87% $\geq$ *5–Somewhat Agree*) and visual source (Q6, 86% $\geq$ *5–Somewhat Agree*) to be accurate given the context of the corresponding conversation.

We analyzed a subset of our visual intent prediction model's results on the test set and qualitatively showcase some examples in Figure 2.7. The model demonstrates good performance and consistency across different topics. It understands the context of a sentence and recommend different visual types, content, and sources based on that context. It also suggests multiple visuals if there are multiple visual intents in a sentence, and can retrieve accurate visuals for ambiguous descriptions. These behaviors were encoded in our collected dataset and training process. In our test dataset, 27 out of 319 sentences have multiple visual suggestions. In future, we would explore training our model on a wider range of visual intents, estimating the importance/appropriateness scores of visuals and allowing users to filter for more or less visual suggestions.

2.6 Visual Captions System

With the fine-tuned visual intent prediction model, we developed Visual Captions on *AR-Chat*, a web-based rapid prototyping platform where developers can quickly build and deploy

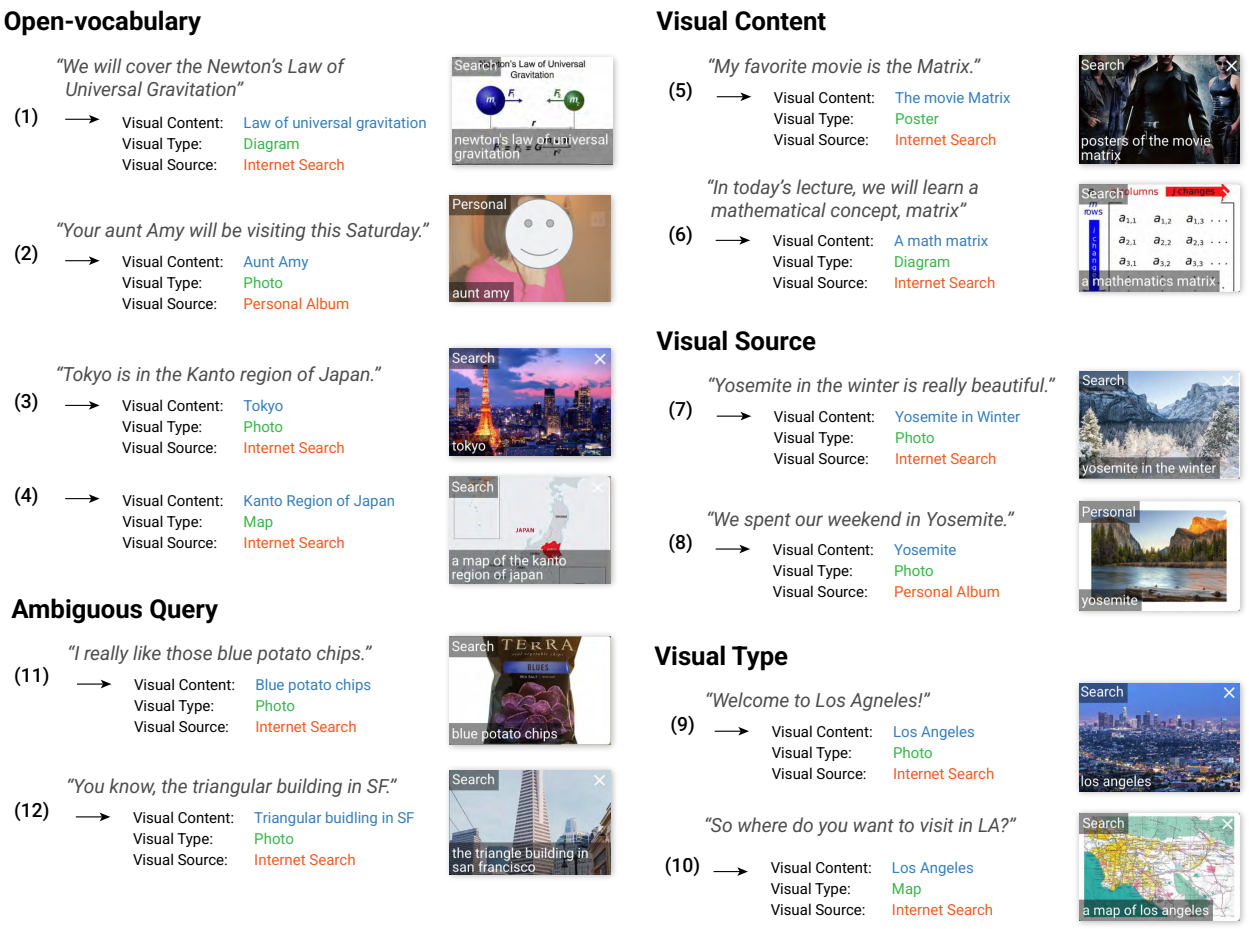


Figure 2.7: Examples visual intent predictions by our model. Compared with keyword-based approaches, our system could handle open-vocabulary conversations and contextually predict visual content, visual source, and visual type.

augmented communication systems. We additionally built a *settings page* for users to customize the visual types to generate, suggestion modes of the AI, and the visual layout.

2.6.1 Visual Captions Interface

2.6.1.1 Generating Visual Suggestions

Visual Captions automatically suggests relevant visuals based on users' conversation content (Figure 2.8). Our system continuously retrieves the automatic captions from Google Meet, and queries for a window of captions every 100 ms. The queried caption is pre-processed and sent as the input to the visual intent prediction model. This query window is customizable on the settings page (2.6.2). By default, our system queries the last two sentences, signified by end-of-sentence punctuation (".", "?", or "!"). To enable responsive visuals for incomplete sentences (*e.g.*, *"Andy Warhol is one of"*), our system also queries visuals if it has more than $n_{\min}$ words ($n_{\min} = 4$ by default).

The model predicts (1) the information to visualize, (2) the type of visual to present, and (3) the source for the visual. For example, it may suggest visualizing *"Santa Monica pier at night"* with an *image* from *online search*. The returned information initiates different pipelines based on the predicted visual type and source. For example, if the model prediction returns *"A photo of me and my family at Disneyland from personal album"*, our system will run a personal album search and return the photo with the highest CLIP score [104], *i.e.*a strong relationship between the language in the text and the visual information in the image. If the model predicts *"a map of Los Angeles from online search"*, our system will run an online image search with the search term *"A map of Los Angeles"*, using the Microsoft Bing Image Search API[5]

Visual Captions then creates a *Visual Widget* object which contains attributes *imgURL* (the URL of the retrieved visual), *description* (the search term), *visual source*, and *visual*

[5]Bing Image Search API: `https://microsoft.com/bing/apis/bing-image-search-api`

type. Widgets are rendered as an HTML element with the visual, the description in the bottom-left corner and the source in the top-left corner (Figure 2.9), and added to the video conference interface. Visual widgets are by default 50% transparent (customizable setting) to make them more ambient and less distracting to the main conversation, and change to non-transparent on hover.

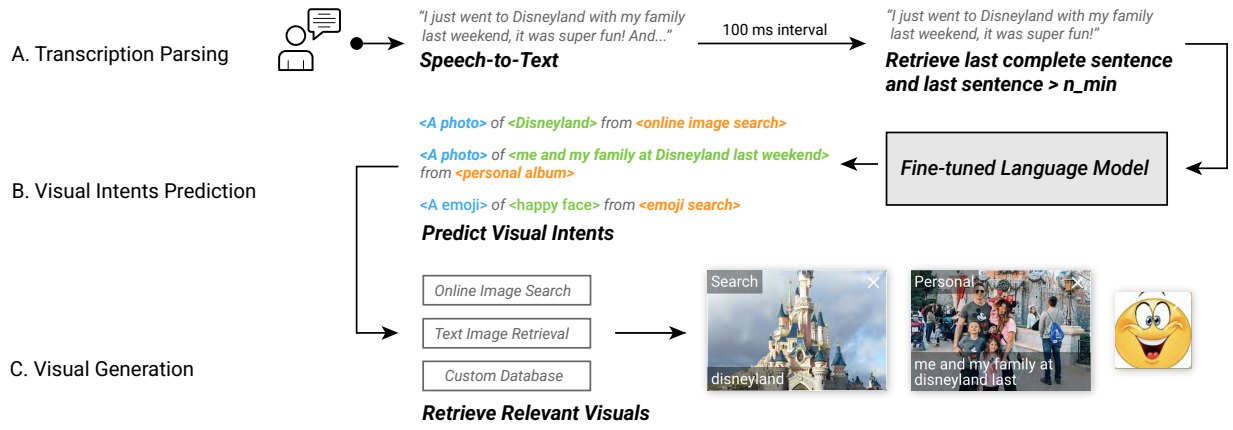


Figure 2.8: System workflow of Visual Captions. The workflow consists of three major steps: A. Transcription Parsing; B. Visual Intents Prediction; C. Visual Generation.

2.6.1.2 Scrolling View

In *auto-suggest* and *on-demand-suggest* modes where users have to explicitly approve our systems' visual suggestions, all generated visual widgets are first displayed in a private *Scrolling View* (Figure 2.9A). The scrolling view automatically updates when new visuals are suggested by the system, and removes the oldest visual widget if it exceeds the maximum amount (customizable *# Max Visuals* in the settings page). Similarly, emojis are displayed in a separate scrolling view in the bottom right corner of the screen (Figure 2.9C). When proactively suggesting visuals, our system on average suggests a new visual every 7.6 seconds in our user studies.

2.6.1.3 Spotlight View

To make an image visible to all parties, the user may click the widget to move it to the *Spotlight View* (Figure 2.9B). Visuals in the spotlight view can be moved, resized, and deleted.

2.6.2 Visual Captions Settings

The Visual Captions settings page allows users to fully customize how they prefer to control and display AI-suggested visuals. For system functionality, users can enable or disable Visual Captions, Emojis, and visuals from personal albums. Users can additionally control what prediction model to use (from "Most capable, but slower" to "Fastest, but less capable"), and how the system queries for visuals (after certain number of words, or after a complete sentence). The system can also suggest visuals based on other meeting participants' speech if enabled (*All Participants' captions*). The layout of Visual Captions on Google Meet is also customizable on the settings page.

2.6.3 AI Proactivity in Visual Captions

Informed by our formative study and pilot study, our system provides three levels of AI proactivity:

Auto-display (high-proactivity)

In the auto-display mode, the system autonomously searches and displays visuals publicly to all meeting participants. AI has full control and no interaction is needed. The scrolling view is disabled.

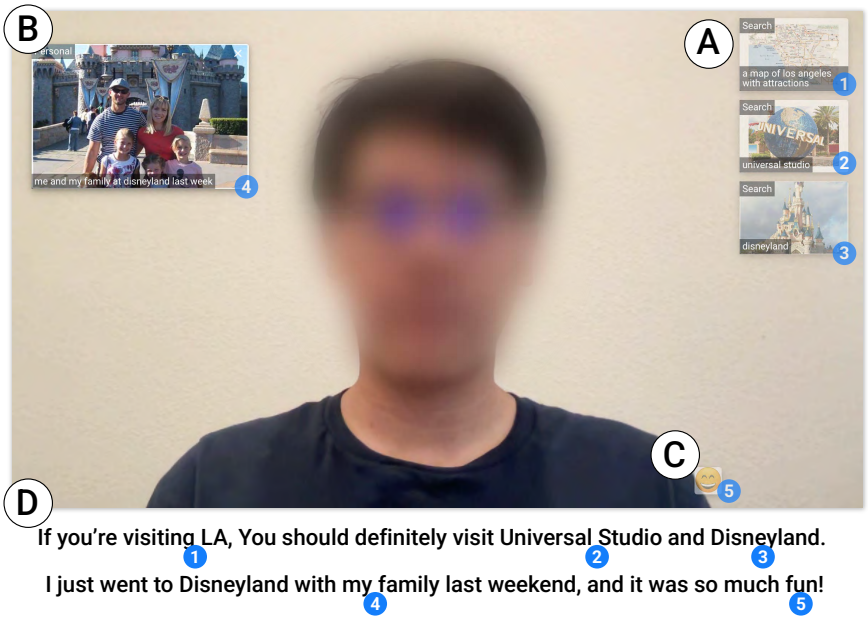


Figure 2.9: In Visual Captions's interface (default auto-suggest mode), the *scrolling view* (A) displays privately candidates of visual suggestions generated by our visual intent prediction model. Emoji suggestions are displayed on the bottom right corner (C). Users can click and display the visual to the *spotlight view* (B) to share it publicly.

Auto-suggest (medium-proactivity)

In the auto-suggest mode, the suggested visuals will be shown in the private scrolling view. A user then click's a visual to display it publicly. In this mode, the AI is proactively recommending visuals, but the user selects when and what to display.

On-demand-suggest (low-proactivity)

In the on-demand-suggest mode, the AI will only suggest visuals if a user taps the space bar. The system immediately queries the captions and stays on for 3 seconds to query the following speech.

2.7 Evaluation

We first conducted a user study with 26 participants to evaluate Visual Captions. We used a mixed-methods study design to gather both data from participants' survey responses and semi-structured interviews. The study examined the following two research questions:

RQ1: How do people use visuals in real-time conversations, and how do visuals affect people's communication?

RQ2: How do people prefer to interact and collaborate with an AI system (Visual Captions) when actively engaged in synchronous human-human activities?

2.7.1 Pilot Study

We conducted three one-hour long pilot studies with 6 participants (3 pairs) to gather initial feedback around the interaction and interface design of Visual Captions. All 6 participants were recruited within Google. In the pilot study, participants (labeled as P1 to P6) were asked to use Visual Captions for a scripted conversation and an open-ended conversation to test out the system, followed with a contextual interview around system's usability.

Participants identified three main issues: the emoji suggestions were distracting and less useful, the lack of an "on-demand" mode, and the lack of customization options. In response, Visual Captions was updated to move the emoji suggestions to a separate, ambient area of the screen and reduce their size, add an "on-demand-suggest" mode which only suggests visuals when the user explicitly requests them, and add a settings page to allow for customization. We integrated all improvements and conducted our formal study with an additional 20 participants.

2.7.2 Materials

We created three scripted dialogues on different topics: discuss where to visit in Los Angeles, order food in a Japanese restaurant, and chat about what they did over the weekend. All conversations were around 5 minutes. Conversations were scripted to provide controls for our data analysis, and allow Visual Captions to suggest a variety of visuals for both participants throughout the study. Participants were also allowed to go off-topic a little bit during the scripted conversations.

2.7.3 Participants

We recruited 20 participants (9 female and 11 male) from our institution's email list and internal communication channel, labeled as P7 to P26. Participants were 21–61 years old (avg=29.5, std=9.4). Participants had varying technical and non-technical backgrounds including students, software engineers, research scientists, designers, and product managers etc. 10 participants use video conferences multiple times per day, five daily, four multiple times per week, and one weekly. Participants were recruited in pairs to fit the one-on-one conversation scenarios in our study. Eight participants were not close friends with their conversation partner. Each participant was compensated with a $25 gift card for their completion of the study.

2.7.4 Procedure

We conducted a one-hour-long study with each participant pair. We set up two meeting rooms in our office. Each meeting room was equipped with a laptop computer connected to a 32-inch monitor, an external keyboard, and a wireless mouse. We created a Google Meet room with Visual Captions connected before the study. We also spent 10 minutes instructing participants how to use Visual Captions and its settings page. Each participant tried all features and experimented with the three AI proactivity levels before continuing to the main part of the study.

Each participant pair was asked to act out four scripted conversations and have a 5-10 minutes open-ended conversation with each other. We divided the study into three parts: (1) Verbal communication vs. communication with visuals. First, participants act out the same scripted conversation two times to directly compare their experience of conversation with and without visuals. All participants used the auto-suggest mode. (2) AI proactivity. Participants then proceed to test the auto-display and on-demand-suggest mode with two additional conversations. (3) Open-ended conversation. Participants use their preferred customized settings to open-endedly chat with their study partner. After each conversation, we recorded participants Task Load Index [53] and Likert scale ratings on their feeling of control, intrusion, error, and interruption when interacting with the system. After the open-ended conversation, participants rated the usefulness of visuals in conversations. Lastly we asked participants to compare their overall experience with and without visuals and different AI proactivity levels through semi-structured interviews. We recorded the audio track for the entire study and the screen portion of using the interface.

2.7.5 Findings: Using Real-time Visuals in Conversations (RQ1)

We present our findings on the impact of visual augmentations in conversations (RQ1) and how people prefer to interact with AIs with different proactivity levels (RQ2).

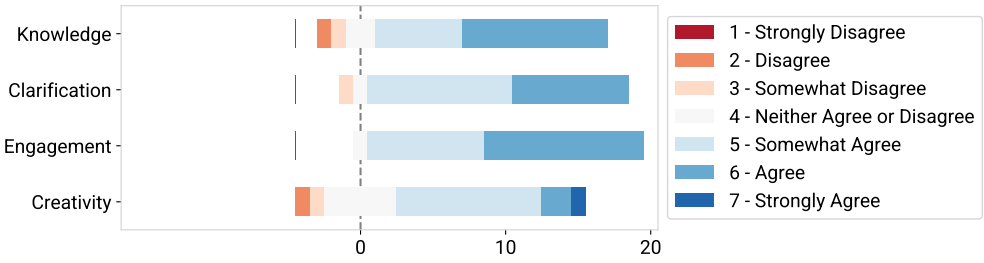


Figure 2.10: Participants' Likert scale ratings to (1) Knowledge: The suggested visuals can help me understand some unfamiliar terms and concepts; (2) Clarification: The suggested visuals can help clarify ambiguity in the conversation; (3) Engagement: The suggested visuals can make the conversation more interesting and enjoyable; (4) Creativity: The suggested visuals provided topics to talk about or led the direction of the conversation.

2.7.5.1 How do people select and display visuals?

Participants in our study exhibited different practices when selecting and displaying visuals. In the auto-suggest mode, some participants only looked at visual suggestions in the scrolling view when they wanted to show something visually. For instance, P9 mostly focused on the conversations and would not notice visuals coming up. For him it was more of a *"purposeful action"*. In contrast, P9's study partner P10 continuously looked for new visuals whenever something popped up, and always *"take a quick look to see if it was something useful to show"*. P15 reported glancing at suggested visuals every time he finished his sentence.

14 participants described that they would click and display a visual to their study partner when they were trying to convey complementary information. P14 and P21 also showed visuals when they were trying to emphasize things. Participants' decision of whether to display a visual or not was largely impacted by the conversational context and social norm. For example, P11 would use more visuals in relaxed and casual conversations as something funny to chat about. P18 said *"it depends on who you're talking to"*. P13 mentioned her hesitation when the system suggested a visual based on a question she asked:

"Sometimes it shows visual answers of what I was asking for, for example 'what is tempura?'. Like should I be the one to show it?" – P13

2.7.5.2 How are visuals useful?

Helps Explain and Understand Unfamiliar Concepts. Participants found real-time visuals to be useful in conversations in multiple aspects. From the study survey responses, participants rated the statement "The suggested visuals can help me understand some unfamiliar terms and concepts and make the conversation more informative" with a *5—Somewhat Agree* or greater 80% of the time (Figure 2.10). They found Visual Captions useful to learn unfamiliar concepts or words without having to ask explicitly:

"The system is especially helpful when it shows me something I don't know, like in this example it shows me a picture of Rodeo Drive. Whenever I'm confused I can just take a look at the right side and see intuitively what they are." – P10

In addition to understanding unfamiliar concepts in conversations, participants found Visual Captions to be helpful in explaining things to their conversational partners. P14 used visuals to more easily describe places to visit to her friend:

"When I would really want visuals is like people don't know what I was talking about. For example when I just mentioned Santa Monica Pier, it's great that I can easily explain what it is." – P14

Makes Known Information More Intuitive 6 participants also reported that visuals helped them with known phrases by making the information more intuitive and easy to convey. For example, P20 enjoyed having visuals of famous paintings to supplement his description: *"When I talked about the Starry Night painting it was really good to have a picture there."* Participants reported that when they wanted to quickly explain a concept without having to describe all details, they found visuals helpful:

> *"Having it pop up and being able to see what the dishes would have been extremely useful. We just had a visual representation of the appetizers, and be like oh I want this one and that one, and then we would see it very quickly."* – P16

Clarifies Ambiguities in Language. Participants reported that Visual Captions helped them clarify ambiguities in their verbal conversations. In several cases, participants used visuals to clarify objects or information that was abstract. 18 participants (Figure 2.10) agreed that the suggested visuals help clarify ambiguity in conversations ($\geq$ *5 — Somewhat Agree*). P21 mentioned that during her conversation, *"whenever you're bringing up something that he wasn't sure which I'm talking about, you can just select the right image."* Similarly P17 described a case where he could select the correct visuals to illustrate what he meant:

> *Back in the beginning when we were talking about the Avatar, there are like four or five different versions we might be discussing, the picture helped crystallize it instantly.* – P17

Makes Conversations More Fun and Engaging. 19 participants (Figure 2.10) found visuals to help make conversations more fun and engaging ($\geq$ *5—Somewhat Agree*). Specifically, participants were able to communicate with more information and *"it makes the conversation longer and more interactive"* (P9). Participants reported that they had richer conversations with their conversation partners:

> *"When chatting with recommended pictures, our interaction has increased. The conversation is getting longer with more content."* – P14

Interestingly, four participants reported that sometimes incorrect visuals suggestions also made their conversations more fun: *"Especially when it suggests funny visuals that were not expected. Sometimes the AI errors were kinda funny."* (P12)

Reduces Efforts in Search for Visuals. Participants found that Visual Captions reduced laborious search and sharing screen when explaining certain concepts: *"It is very interesting*

and necessary to have them. If we talk about something, we search and this is much faster and easier" (P11). Two participants mentioned they could grant the system access to their personal albums and Instagram posts when talking to their parents or close friends. P20 said it would be *"so much easier for me to show photos from my recent trips when talking about it"*.

2.7.5.3 How visuals impact the way people communicate?

Visual Captions Guided Conversations and Provided New Topics. In open-ended conversations, participants found that Visual Captions guided their conversations as they would often start talking about the visuals when they showed up. For example, P13 described:

> *"I feel like we were responding to the photos. When we were talking about a whale watching tour, it suggested an image of people on a very small boat. We got to further discuss what boat I was on in the tour and so on."* – P13

Visuals are also helpful to find new topics to get to know each other, especially for those who were unfamiliar with their study partners. 13 participants rated the statement "The suggested visuals provided topics to talk about or led the direction of the conversation" with greater or equal to *5—Somewhat Agree* (Figure 2.10). Moreover, in our scripted conversations, a majority of the participants did not strictly follow the script and sparked emotional remarks or engaging actions. For example, after P11 said *"Let me show you a picture of Rodeo Drive"*, P12 answered *"Wow, it appears looking really pretty"* instead of the *"I'll have to check it out"* as originally scripted.

Distractions of Visual Captions. To our surprise, participants did not find conversations with extra visuals to be more mentally (Mann-Whitney U test $p > 0.05$) or physically demanding ($p > 0.05$) than normal conversations when the system is proactively suggesting visuals (*i.e.*auto-display and auto-suggest), and did not find it to be more stressful or

annoying ($p > 0.05$) in any condition (Figure 2.11). While many participants found Visual Captions to be *"not distracting at all"* (P21), some found Visual Captions to disrupt the conversation flow in two ways: interaction required for selecting and displaying visuals (more discussion in subsection 2.7.6), and deciding what visuals to display. Participants reported that sometimes they needed to switch their mind to decide whether a visual is appropriate:

> *"It takes some time to identify which images are proper to share. It interrupts the conversation a little bit, but otherwise I feel OK."* – P7

2.7.5.4 Improvements and Opportunities

Over half of the participants envisioned to use Visual Captions with augmented reality glasses. Participants suggested that it would be more natural and less intrusive, especially for in-person conversations. P12 particularly mentioned the use of hand gestures to intuitively select and display visuals. Other participants also mentioned accessibility use cases of Visual Captions for people with dyslexia or language barriers. In addition, many participants were excited about expanding this tool into professional settings, and use it in their work to retrieve slides, profiles, internal links etc.

2.7.6 Findings: Interacting with AI (RQ2)

2.7.6.1 People have a large variance of AI proactivity level preferred

Participants in our user study had diverging preferences over the AI's proactivity level, even within the same study pairs. Six participants preferred the *auto-display* mode, seven preferred the *auto-suggest* mode, and seven preferred the *on-demand-suggest* mode.

Auto-Display Mode. Participants who preferred the auto-display mode liked that almost no interaction was needed to activate and display the visuals, as P7 mentioned, *"Not having to click is huge for me."* On a scale from 1 to 7, participants rated their effort using the

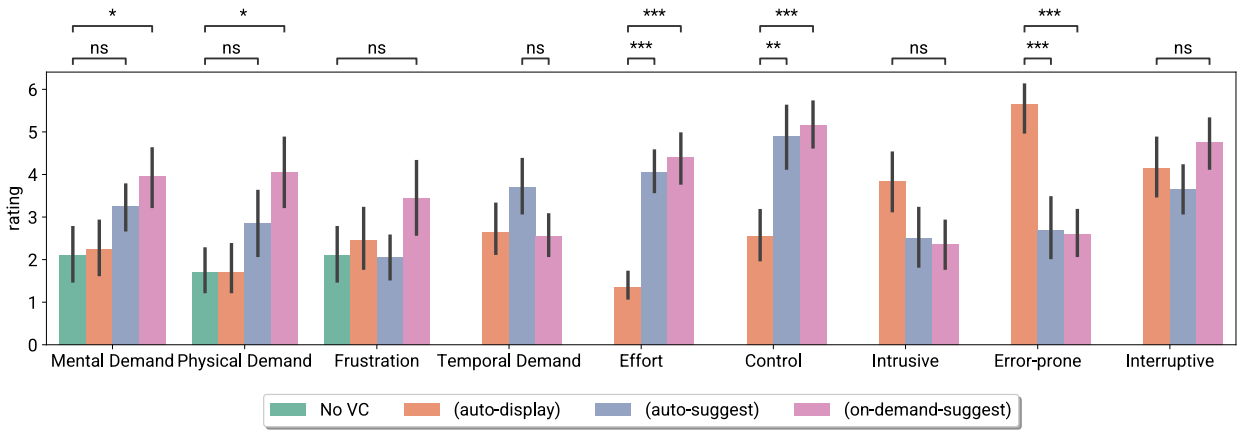


Figure 2.11: Participants' Task Load Index and Likert scale ratings (from 1 - Strongly Disagree to 7 - Strongly Agree) to four conversations with different Visual Captions modes: No VC, Auto-Display, Auto-Suggest and On-Demand Suggest, with 95% confidence interval bar. In the Task Load Index questions we asked participants their feeling of mental demand, physical demand, frustration (how stressed and annoyed), temporal demand (how hurried and rushed), and effort. In Likert scale questions we asked participants about their experience interacting with the AI, and their feeling of having control, AI intrusiveness, difficulty to override errors, and AI's interruption to the conversation. We additionally annotated statistical tests (Mann-Whitney U) results between different conditions: ns (non-significant), * ($p<.05$), ** ($p<.01$), *** ($p<.001$).

auto-display mode ($\mu = 1.35, \sigma = 0.73$) to be less than auto-suggest ($\mu = 4.05, \sigma = 1.07$) and on-demand-suggest modes ($\mu = 4.4, \sigma = 1.28$), with statistical significance (Mann-Whitney U test $p < 0.0001$).

However, some participants complained about it being prone to errors and having less control: *"Sometimes I don't want to emphasize the point, but it suddenly shows an image..."* (P9). From our survey responses, participants felt that the auto-display mode provided significantly less ($p < 0.01$) control ($\mu = 2.55, \sigma = 1.32$) compared to auto-suggest ($\mu = 4.9, \sigma = 1.67$) and on-demand-suggest ($\mu = 5.15, \sigma = 1.28$), and were harder to override system mistakes ($p < 0.001$). For example, P17 had to pay attention all the time to not risk showing something inappropriate: *"It interrupt my conversation because I always pay attention to the spotlight view, and immediately check whether the image displayed is appropriate."* In contrast, participants who preferred auto-display felt that the incorrect visuals did not matter that much:

> *"If it shows an image incorrectly, I just feel like it's an imperfect tool. Just like textual captions, it's not accurate and I just accept that fact."* – P20

Privacy is another concern in the auto-display mode. Although we did not observe any statistical difference in participants' responses, some participants mentioned that the system could accidentally show private photos without permission: *"I don't want to show my personal album without confirmation."* (P7) In our user study, while our system can suggest personal visual intents, we did not use real images from personal albums for privacy reasons. This is a limitation of the study, and it would be important to understand how people would feel about having real personal visuals suggested or displayed, especially in the auto-display mode. In future, implementing safeguards such as an extra confirmation of private photos may give users more control over their privacy.

Auto-Suggest Mode. Seven participants preferred the *auto-suggest* mode, as they felt that they could better understand the system's capabilities in the private scrolling view and

share with others when needed: *"During the conversation, I really like talking when these things pop out, so we really know what it is like"* (P8). Participants who preferred this mode usually *"only look for visuals when I want to show something."* (P12).

Other participants did not like selecting from the scrolling view: *"It takes some time to identify which images are proper to share."* (P22). Some mentioned that this mode increased the cognitive burden and distracted their attention as visuals are being constantly suggested in the scrolling view:

> *"For me it was kind of a distraction from the the conversation because typically I don't look away from the person when I talk, and I think having something moving in the frame is kind of distracting"* – P14

On-Demand-Suggest Mode. Although participants rated the on-demand-suggest mode to the most mentally and physically demanding (Figure 2.11), seven participants preferred this mode. Participants considered it to be the least distracting and appreciated having full control over when to query for a visual:

> *"I prefer the tap one since I feel like I was controlling whether I want to show something. It's less mental overload and distraction because I would only activate it when I want, and <it> also <has> less things on the screen."* – P13

P17 also found this mode to seemingly provide the most accurate visual suggestions, because *"I'm only seeing visuals I requested."*

Participants who did not like on-demand-suggest felt too much interaction effort was needed to display a visual. Interestingly, although the latency of visual suggestions should be the same for all three modes on average, participants repeatedly reported that they felt significantly more delay when using the on-demand-suggest mode. We hypothesize that when humans are taking the initiative, people are more sensitive to when they started the query, and thus a more obvious feeling of delay. While with proactive AI, people are mostly paying

attention to the main conversation. Because of this perceived delay, several participants found it hard to understand the internal state of the system: *"It's hard for me to know when to press the spacebar, and once I pressed the spacebar, did it trigger the system or there's just no suggestions?"* (P8). Participants sometimes had to interrupt their conversation to wait for something to show up:

> *"I have to pause for a bit after I hit the space and hoping that the right thing will show up. "* – P21

2.8 Case Study: Usage of Visual Captions In-the-wild

To further understand how people would integrate real-time visual augmentations into their daily conversations, we distributed Visual Captions let people try it in-the-wild for two weeks. We recruited 10 participants (5 females, age: $22 - 33$, avg=28.5, std=3.8) via email invitations and social channels in Google, labeled as U1 to U10. Six users previously participated in our lab study. We encouraged participants to use Visual Captions in every meeting during the week. Participants filled in an entry survey and an experience sampling survey every three days.

Using Visual Captions Beyond One-on-one Conversations. One limitation of our case study is we recruited participants within our institution due to confidentiality, and participants had mostly work-related meetings where Visual Captions was not specifically designed for. However, all participants still used Visual Captions in various scenarios: 4 users used Visual Captions in multiple meetings per day, 2 users about once a day, and 4 users multiple times a week.

Participants reported not only using Visual Captions in one-on-one meetings, but also in team meetings, remote social events, and listening to online talks, where multiple parties were involved. *"I used Visual Caption in a team social with my colleagues for 45 minutes, we were talking about favorite board games and weekend plans."* (U7), *"I used VC with other*

team members to talk about their experiences with a demo. Emotions during the talk were conveyed by emojis." (U3). In such many-to-many scenarios, participants found Visual Captions especially attract attention from the audience:

> *"I was doing a self-introduction in a group social event, and it really attracted people's attention and increased the fun and engagement at the beginning."* – U2

Visual Captions Improve Expressiveness and Reduce Social Awkwardness. Some participants used Visual Captions as an implicit self-expressive tool for listeners. They noted that Visual Captions was especially helpful in large team meetings, as it allowed them to express their thoughts and feelings without interrupting others, through a parallel visual channel. *"Showing the suggested emoji on the screen also prevents interrupting with other people's talking while expressing oneself's agreement or rejection."* (U5)

In addition, Visual Captions reduces embarrassment and frictions for foreigners or laymen of specific topics in social scenarios. For non-native speakers, Visual Captions offers a parallel channels for users to quickly catch up with concepts that are familiar to others. "*It helped understand unfamiliar words in English as a non-native speaker. E.g., Andromeda.*" (U3) "*'VC pops up images for words that I don't understand, like 'groomhaven', 'borg sphere' in a social meetup."* (U9)

Users Prefer Different AI Proactivity in Different Scenarios. All six participants who participated in our lab study reported continuing using their preferred AI proactivity level in the case study. However, many participants reported that during the deployment, they were more likely to change the AI proactivity levels depending on the type of meetings, indicating their preferences for different levels of AI proactivity in different social scenarios. For example:

> *"I use automatic (auto-display) mode all the time, but change to on-demand mode in important meetings because I don't want to interrupt other speakers."* – U6

In one-on-one meetings, users may prefer a higher level of AI proactivity to enhance their own expressiveness, while in important group meetings they may prefer a lower level of AI proactivity to avoid interrupting other speakers. Similarly, U7 described that *"I used VC to show images of my own words in 1:1 meetings but show images of everyone's words in a group meeting using the auto-suggest mode."* This highlight the importance of providing users with flexible control over the level of AI proactivity in Visual Captions.

2.9 Discussion and Future Work

We describe the limitations of our system, discuss the implications and envision future opportunities:

Visual Appropriateness in Conversation. We identified two major causes of errors in Visual Captions. First, we observed that many of the visual suggestion errors come from incorrect captions. As our system relies on automatic speech recognition (ASR) as the input to the model, imprecise ASR results may lead to drastically incorrect visual suggestions. In future, we could leverage methods like topic modeling [137] and summarization [49] to compute a *visual appropriateness score*, to determine how relevant the visual is to the current conversation.

Several participants also mentioned that one of their main sources of distraction was to think about whether they want to display the visual. To this end, another aspect of the visual appropriateness measure would be to understand how appropriate it is to show the visual to others, given the current context, relationship with the person, and timing. Having a threshold to filter out all inappropriate visuals would be especially helpful in the auto-display mode.

Promptness-Accuracy Trade-off. The second main cause of error in Visual Captions is that it greedily retrieve unfinished sentences to ensure *promptness* in visual suggestions. For example, in the sentence *"San Francisco is a beautiful city"*, if the algorithm acts too

aggressive it may incorrectly query *"San"*, leading to incorrect images. This problem of *when to query for visuals* is fundamentally similar to the problem of *when to transcribe* and *when to translate* in live speech. Recent research have proposed the Wait-K policy and to predict stability scores of new phrases [7]. In future we could employ a similar approach to search for visuals that have high probability to be stable.

Improvement of Visual Suggestions.

Personalized Visual Suggestions: Many participants use Visual Captions to explain information that their study partner did not know, and understand concepts that they were not familiar with. However, already known concepts may feel redundant. It still remains an open question that how to model people's familiarity with different concepts or knowledge, without privacy risks. In our case, we could explore online learning algorithms to continuously adapt the model to user's personal preferences.

Integrating Text-to-image Models: Recent advances in contrastive language models [104] and diffusion models in vision [73, 122, 57, 116] have fostered a few groundbreaking text-to-image systems: DALLE-2 [105], ImageGen [115], Parti [153], etc. While such models generate delicate visuals based on descriptions, our system infers the implicit visual intent of human conversations. However, it is interesting to explore integrating these text-to-image models into our system when people's visual intents are detected to be creative and imaginative, in scenarios such as brainstorming and storytelling.

Visual Coherence: It would also be beneficial for Visual Captions to consider coherence in visual styles and formats throughout a conversation. For example when discussing different data plots, it would be beneficial if the plots generated throughout the conversation were coherent in style, such as being generated with the same visualization tool like pyplot. In future, we plan to explore the use of consistent tools and style metrics to generate visuals that are coherent across a conversation.

The Future of Augmenting Verbal Communication with Visuals. In this paper, we

explored augmenting verbal communication with visuals on videoconferencing platforms, as it is one of the most common way people communicate. However, as shown in our design space (subsection 2.3.2), we believe that the possible applications are much broader. For example, future systems could explore augmenting in-person conversations with 3D visuals and hand gestures in augmented reality or a metaverse like Geollery [37, 38], generating visual effects from verbal speech, or using real-time visuals for brainstorming and to assist creative writings.

Does a One-size-fits-all Human-AI Interaction Mode Exist? Participants had a large variance in the level of AI proactivity they preferred — 6 preferred auto-display, 7 preferred auto-suggest and 7 preferred on-demand-suggest. Participants also would like to use different AI proactivity levels under different social scenarios. While a number of recent research have proposed new human-AI collaborative systems and aimed to find an optimal human-AI interaction paradigm, there is an opportunity to investigate how to design and provide systems that are adaptive, with different levels of human and AI initiatives.

2.10 Conclusion

In this work, we presented our vision and implementation of Visual Captions, a human-AI system designed to visually augment real-time conversations. We shared how we worked with crowdworkers to gather 1595 visual intents to fine-tune a large language model to power our system, and report on the results from a formative study that informed the development of a design space for systems like ours. Our system, Visual Captions, was designed as a virtual camera and is therefore compatible with today's popular video conferencing systems. As a Chrome browser plugin, it can leverage state-of-the-art real-time speech transcription and be easily installed by end users with no other dependencies. This approach allowed us to evaluate the capabilities in both a user study and a longer deployment study. The results suggest that Visual Captions has the potential to facilitate live conversations with valuable

visual content.

CHAPTER 3

From Reactive to Proactive: Conversational Agents with Inner Thoughts

If Chapter 2 expands interaction beyond a single linear channel, Chapter 3 tackles the second constraint of turn-taking: *reactivity*. In most conversational systems, the AI is designed to wait for a prompt or to speak only when addressed, implicitly treating initiative as belonging to the human user. However, in real multi-party conversations, participation is not only a matter of detecting external cues; people continuously form covert thoughts while listening, and later choose to contribute when motivation and timing align.

Inner Thoughts reframes proactive AI as a process driven by an internal stream rather than by turn-taking prediction. Instead of merely reacting to observed cues, the framework equips an AI participant with a continuous, covert train of thoughts that runs in parallel with the overt conversation and uses an explicit evaluation of intrinsic motivation to decide whether and when to speak. This perspective is also motivated by the limits of next-speaker prediction approaches, which often struggle when turns are not explicitly allocated and when participation depends on intrinsic factors rather than observable context. Chapter 3 therefore connects directly to the thesicoms claim that moving beyond turn-taking requires rethinking AI as an agent capable of self-initiated contribution, not only prompt-following response.

3.1 Introduction

Recent advances in Large Language Models (LLMs) have demonstrated their ability to generate high-quality text in response to human input, finding application in areas ranging from Q&A systems to writing assistants. Yet, most current LLM-based systems treat AI as passive respondents, responding only to explicit human prompts. Imagine a scenario where people are planning a trip with an AI agent: they must constantly prompt the AI, which passively waits for instructions instead of actively contributing. On the other end of the spectrum, systems like GitHub Copilot[1] tend to overcompensate, offering constant suggestions that can overwhelm users. Neither extreme — AI that is only reactive nor AI that is always responding — is ideal.

In the context of conversations, a *proactive* AI agent should be able to autonomously participate in socially appropriate moments, providing relevant input without requiring explicit cues. This is particularly challenging in multi-party conversations. Dyadic human-AI interactions (*e.g.*using Siri) often predict turn-taking based on speech features such as pause or stop words, and the next turn will be automatically allocated to the other party [121, 27]. However, in multi-party settings, these cues could be ambiguous, and multiple possible speakers may take the floor. Repeatedly prompting AI during group interactions can also become cumbersome and can disrupt the natural flow of the conversation, as illustrated in the example of trip planning.

Previous systems typically first predict the next speaker (*i.e.*turn-taking prediction) and then generate the next response based on conversational and contextual information. For instance, some approaches rely on the last few turns of conversations to predict the subsequent speaker [142, 34, 42], while others utilize multimodal cues, such as eye gaze and non-verbal signals [13, 14, 15]. Despite these efforts, on turn-taking prediction, they still fall short and struggle to beat the simple "repeat last" baseline strategy in social conversation

[1]https://github.com/features/copilot

contexts [142, 34]. Our formative evaluation (Table 3.1) also shows that when it comes to predicting the next speaker, fine-tuned LLMs perform no better than random guessing unless the next speaker is allocated (*e.g."What do you think, Alice?"*). In addition, after determining the next speaker, existing works tend to use predefined speaker personas [155, 150] as additional input to guide response generation, or expand persona with commonsense [69]. However, these additional inputs and profiles are fixed and static during conversations, instead of changing through time as humans did.

We suggest an alternative and reversed perspective to think about AI proactivity: Consider how humans chat about what we did over the weekend. As we listen to others speak, we process their words, reflect on our experiences, and develop an internal train of thoughts — cognitive psychologists highlight this as the distinction between *covert responses* (internal thoughts and feelings) and *overt responses* (verbal utterances or gestures) [41, 112] in the human communication process. Then, at some point, we may feel a strong urge to share our thoughts. This might happen when we seek clarification or when someone mentions an activity we also participated in, sparking our desire to contribute. With this intention in mind, we then look for a socially appropriate moment to participate.

In this paper, we propose a new approach to proactive AI in the context of multi-party, text-based conversations: rather than simply predicting conversational turns, we explore proactive AI driven by its own internal "thoughts". We introduce the **Inner Thoughts** framework. Inspired by cognitive architectures and LLM prompting techniques, this framework comprises five stages: *trigger, retrieval, thought formation, evaluation, and participation*, which enable AI to continuously generate a train of thoughts in parallel with an ongoing conversation, utilizing both long-term and working memory. The AI participant then determines whether to engage in the conversation based on an evaluation of its *intrinsic motivation* to express a particular thought at that moment.

To model intrinsic motivation, we conducted a think-aloud study with 24 participants, each of whom participated in four synchronous, text-based online group chats. Using the

affinity diagram approach, we organized and analyzed the interview notes, and derived 10 high-level themes on how individuals decide to engage in conversations. These heuristics were then formalized into automatic evaluation criteria (*e.g.*relevance, information gap, etc.) for AIs to quantitatively rate their intrinsic motivation to participate.

We implemented our framework as two systems: a multi-agent playground web app and a chatbot. Our technical evaluation shows that conversational agents driven by Inner Thoughts significantly outperformed a next-speaker prediction plus persona baseline across all seven evaluation metrics, including turn appropriateness, coherence, anthropomorphism, perceived engagement, intelligence, initiative and adaptability. Participants preferred the Inner Thoughts approach over 82% of the times, noting more natural turn-taking and contextually aware contributions, while the baseline was less preferred for its mechanical and disjointed responses.

In summary, we contribute:

- Inner Thoughts, a framework for enabling proactive conversational AI by creating a parallel train of thoughts and modeling its intrinsic motivation to express these thoughts.
- Heuristics derived from a study with 24 participants that reveal how humans choose to express or hold back their thoughts during conversations. These heuristics are instantiated as evaluation metrics for modeling AI's intrinsic motivation to participate.
- Two implementations of the Inner Thoughts framework: a multi-agent simulation playground web app and a chatbot (named *Swimmy*[2]), both deployed and open-sourced at `https://liubruce.me/inner_thoughts/`.
- A technical evaluation and user study comparing Inner Thoughts with baseline models.

[2]We named our chatbot *Swimmy* based on a quote by Edsger W. Dijkstra: *"The question of whether a computer can think is no more interesting than the question of whether a submarine can swim".*

3.2 Related Work

Our work builds on previous research in proactive conversational agents, turn-taking in multi-party conversations, and thought-augmented LLMs.

3.2.1 Proactive AI and Conversational Agents

Proactive AI dates back to earlier work on mixed-initiative interaction [3, 60]. In contrast to AI that only passively responds to human queries, mixed-initiative interaction envisions agents that autonomously understand when to take what action, such as the LookOut system [60] that automatically identifies related dates and events in emails and then proactively suggests them to users as calendar events. In 1996, Rhodes et al. [108] introduced one of the pioneering systems to continuously supply relevant information through observation of human activities. Andolina et al. [5] developed SearchBot, which offers ongoing suggestions of related documents and entities unobtrusively [4] during voice interactions.

While proactivity is a recurring theme in conversational AI research, most proactive conversational AIs focus on task-oriented contexts [81, 47, 82], with the aim of helping users achieve specific objectives. Social conversations, which can expand on open topics without having any goal to complete, are rarely addressed. In addition, past research tends to focus on generating proactive text responses to help lead and guide the conversation [35], for example, the ability of "learning to ask" [12, 34, 136, 107], understanding and initiating topic shifts [84, 129, 149], and planning future conversation [72, 93, 132] etc.

In this paper, we focus on investigating how to enable AI to *proactively engage* in multi-party conversations: how AI can determine the appropriate moments to speak and what contributions to make. We also choose to investigate *social* conversations where unlike task-oriented dialogue, the objectives are often ambiguous, and the actions required from the AI are not clearly defined.

3.2.2 Turn-taking in Multi-Party Conversations

For a conversational agent to engage proactively, it must understand and manage turn-taking, deciding who should speak at the end of each turn. Modeling turn-taking is still an area of active research. Existing approaches often employ an explicit mechanism, such as a "send" button [143], push-to-talk [55, 133], and wake-words (*e.g.*"Hey Siri") [74]. However, the use of explicit cues can be viewed as less conversational from users' perspectives [145]. Mainstream conversational AI systems also use silence to detect the end of a user's turn. However, studies show pauses within turns are typically longer than gaps between turns in human conversations [17, 130], making silence an unreliable cue for turn-taking. More importantly, this method does not generalize to multi-party conversations. In dyadic interaction, it is always clear who is supposed to speak next when the turn is yielded [121]. In the multi-party case, this becomes more ambiguous since there is more than one potential speaker who might take the turn.

Beyond an explicit mechanism, machine learning researchers have proposed data-driven methods to manage turn-taking in these conversations, primarily leveraging conversation history to predict the next speaker (*i.e.*the *next-speaker prediction* task) [142, 34, 42]. However, these methods have shown limited success. Notably, they have often failed to outperform the simple "repeat last" baseline strategy in social conversation contexts [142, 34]. In addition to using only textual data, research in HCI and HRI have leveraged other contextual, non-verbal information and "turn-taking cues", for instance, eye gaze (*e.g.*looking at addressee) [101, 102], breathing (*e.g.*breathe in and out) [65, 89], prosody (*e.g.*rising or falling of pitch) [44, 39, 40, 85] and the status of the human user (*e.g.*passing by, stopping) [13, 14, 15] to decide if an AI should engage at a certain moment of the conversation or not.

Previous approaches on mediating turn-taking often relied on conversation history and contextual information, and typically treat the AI as a reactive agent. Inspired by human

behavior, our Inner Thoughts framework takes a different perspective by modeling intrinsic motivation to speak.

3.2.3 Language, Thought, and LLM Agents

Recent advances in large language models (LLMs) have incorporated intermediate reasoning steps to enhance performance in complex tasks, such as Chain-of-Thought (CoT) prompting [141] whereby LLMs think step-by-step to effectively break down larger problems into reasoning steps, and Tree of Thoughts (ToT) [151] whereby LLMs explore multiple possibilities at each reasoning stage. In addition, self-reflection mechanisms can iteratively improve the model's reasoning. ReAct [152], for example, synergizes reasoning with action-taking by having the model alternate between generating reasoning traces and performing task-specific actions. Reflexion [120] builds on this by equipping models with dynamic memory and self-criticism capabilities, allowing them to refine future actions based on past performance. Expanding on this, Generative Agents [99] simulate human-like behavior by combining memory, planning, and reflection. The recent OpenAI's o1 preview [98] introduces another perspective on reasoning transparency by explicitly surfacing intermediate reasoning steps to make the AI's decision-making process more interpretable to users.

The Inner Thoughts framework we propose diverges from these approaches by simulating an ongoing, parallel stream of internal thoughts that mirror human covert responses. Unlike methods such as CoT, ToT, or OpenAI o1 preview, which emphasize externalizing intermediate steps for reasoning tasks, Inner Thoughts explore leveraging these covert thoughts to equip AIs with the ability to self-initiate actions and engage proactively.

3.3 Next-Speaker Prediction Is Insufficient to Enable Proactive Conversational AI

In this section, we investigate the limitations of the commonly used "next-speaker prediction" strategy [142, 34, 42], and further motivate the need of Inner Thoughts to enable proactive AI engagement in multi-party conversations. While next-speaker prediction perform well when explicit turn-allocation cues are present, we demonstrate that they fall short in *self-selection* cases, where turn-taking decisions are mostly spontaneous and influenced by covert, intrinsic factors of the conversational parties rather than observable contextual cues. Building on Sacks et al.'s *Simplest Systematics* [114], turn-taking in conversations is governed by a set of rules:

1. **Turn-allocation**: The current speaker may select the next, often using cues like gaze or address terms (*e.g."What about you, Alice?"*).
2. **Self-selection**: If the current speaker does not select a next speaker (*e.g."I went to Disneyland last weekend."*), then any party can self-select to take the floor. The first to start gains the turn.
3. If no other party self-selects, the current speaker may continue.

Our intuition is that decisions to self-select and participate are largely influenced by covert internal processes — such as a participant's interest, relevance, or motivation to engage — which are not easily observable from explicit conversational data. Thus, we argue that training machine learning models on next-speaker prediction tasks based on conversation history is inherently ill-suited for self-selection scenarios, because there is no deterministic mapping between prior utterances and the next speaker. To further verify our hypotheses, we evaluate the performance of several Generative Pre-trained Transformer (GPT) variants in predicting the next speaker in multi-party conversations in both turn-allocation and self-selection scenarios.

3.3.1 Hypotheses

We hypothesized that GPT would perform well in turn-allocation scenarios, as these are often signaled by explicit language patterns. However, we anticipated lower accuracy in predicting the next speaker in self-selection scenarios, as these decisions are likely influenced by participants' intrinsic motivations, which are not directly observable from conversational context.

We further expected that fine-tuned models would underperform, especially in self-selection scenarios, as fine-tuning on datasets with high variability in self-selection decisions could introduce noise or misleading patterns.

3.3.2 Materials & Methods

We used the Multiparty Chat Corpus (MPC) [119], a dataset designed to capture social dynamics in multi-party conversations. The dataset includes chat logs from sessions that began as free-flowing and became increasingly structured over time. A key feature of the MPC dataset is the communicative links annotation `link_to`, which identifies whether each utterance was addressed to a specific participant. For our analysis, turn-allocation refers to utterances addressed to a specific individual, while self-selection refers to instances open to all participants (`all_users` in MPC).

The MPC dataset reveals a significant imbalance between these two turn-taking strategies. Out of the total utterances, 95% were self-selection, while only about 5% were instances of turn-allocation. Baseline accuracy for predicting the next speaker in this context was approximately 12.7% (average $\frac{1}{n}$ of all conversations).

We tested six different models. We first evaluated prompting the base GPT-3.5, GPT-4-turbo, and GPT-4o models (#1, 2, 3 in Table 3.1) to predict the next speaker. The prompt first specify the number of speakers in the conversation and their names, and provides the last five utterances from the conversation (following the prediction window configuration

of [34]). The model is instructed to predict the most likely next speaker by name. We also tested zero-shot chain-of-thought (CoT) (model #4) where we prompt the model to provide reasoning for its prediction first. Finally, we experimented fine-tuning GPT-3.5 on the MPC corpus. Using the communicative links annotation (`link_to`), we labeled each utterance based on the participant it was addressed to (model #5) or open to all (self-selection, *anyone* in model #6). We used the same prompt structure as models #1, 2, 3. To create the fine-tuning dataset, we used the MPC corpus and split the data into 70% training and 30% testing sets using a random selection of files. We balanced both sets by including all instances labeled as turn-allocation and randomly sampling an equal number of self-selection instances. Complete prompts used for evaluation are listed in Supplementary Material.

#	*Model*	*Overall*	*Self*	*Alloc*
1	GPT-3.5	0.165	0.066	0.248
2	GPT-4-turbo	0.390	0.099	0.633
3	GPT-4o	0.435	0.121	0.697
4	GPT-4o CoT	0.430	0.187	0.633
5	Fine-tuned GPT-3.5	0.265	0.156	0.378
6	Fine-tuned GPT-3.5 (Speaker name or *anyone*)	**0.810**	**0.853**	**0.765**

Table 3.1: Next speaker prediction accuracy for different GPT models, grouped by the turn-taking type of the current utterance: self-selection (*Self*), turn-allocation (*Alloc*) and overall. The random guessing baseline accuracy is 0.127.

3.3.3 Results

Our results (Table 3.1) demonstrate that all models performed better in turn-allocation scenarios, with consistently higher accuracy, and GPT3.5 performs significantly worse than GPT4 models. In contrast, performance in self-selection scenarios hovered around random chance, supporting the hypothesis that self-selection might be influenced by internal factors that are not easily inferable from the conversational context alone. GPT-4 with CoT reasoning improved predictions in self-selection scenarios but still significantly worse than turn-allocation predictions. As expected, fine-tuned models introduced overfitting particularly in self-selection cases, where the model may have learned patterns of the next speaker that are not truly generalizable.

These findings suggest that context from previous utterances and speaker information is insufficient for accurately predicting the next speaker or determining who should proactively engage, especially in self-selection scenarios.

3.3.4 New Task: Speaker Name or "Anyone"

Given the inherent challenges of predicting specific next speakers in self-selection scenarios as discussed earlier, we additionally experimented modifying the labeling schema to better align with the turn-taking mechanisms in multi-party conversations. In this task, next speakers in turn-allocation scenarios remain labeled with their respective names, while self-selection scenarios are relabeled as *"anyone"* (model #6). This labeling schema results in significant performance improvements in *both* turn-allocation and self-selection scenarios (Table 3.1). Specifically, fine-tuned GPT-3.5 model with this modification achieves an accuracy increase in turn-allocation cases, from 37.8% to 76.5%. This shows that removing the requirement to predict specific speakers in self-selection scenarios eliminates a major source of noise and unpredictability, and enhances model performance across the board.

3.4 Retrospective Think-aloud Study

Findings from section 3.3 show that predicting next speakers in self-selection scenarios requires more than analyzing previous utterances — it hinges on understanding the intrinsic motivations of participants. We are motivated to introduce the concept of inner thoughts for agents and investigate what factors contribute to one's intrinsic motivation to participate. If we are to design a proactive agent system, how should we model its intrinsic decision-making process? Given the agent's thoughts, what factors beyond prior utterances influence its decision to participate? To answer these questions, we conducted a retrospective think-aloud study [52] to observe how human participants decide whether to engage in a multi-party conversation. Specifically, *what factors influence their choice to express or withhold a thought, particularly when the opportunity to speak is open to all?*

3.4.1 Participants

We recruited 24 participants (10 female, 14 male) from our institution in groups of three. Before the study, participants completed the Big-5 personality test [110] and rated their familiarity with one another on a Likert scale (1–7). Participants reported varied levels of extroversion (Max: 97, Min: 2, Avg.: 50.0, SD: 32.3), and most were relatively familiar with one another (Avg.: 5.76, SD: 1.02).

3.4.2 Procedure

Each group engaged in four 10-minute synchronous text conversations on Slack. The conversations covered four topics (trip planning, casual chat, friendly debate, and brainstorming), and participants were free to direct the conversation as they wanted. After each conversation, participants reviewed the discussion utterance-by-utterance, reflecting on their thoughts at moments when they considered contributing or chose to remain silent. We prompted them with the following questions: (1) *What*: What were you thinking? Did you want to say it?

(2) *Why*: Why did you feel the need to say or not say it? (3) *When*: Did you decide to jump in immediately, wait for a pause, or wait for a particular statement?

After the think-aloud sessions, we conducted semi-structured interviews to further reflect on instances when participants felt strongly about contributing or chose to remain silent despite having thoughts to share. Each participant was compensated 14 USD in local currency.

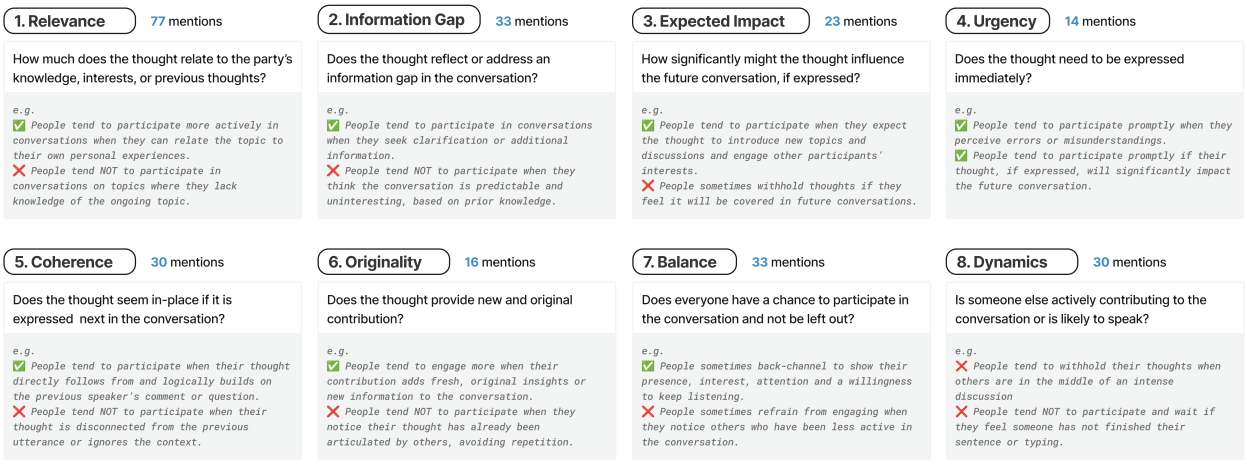


Figure 3.1: People's intrinsic motivation to engage in conversations: Heuristics of what factors influence people's decisions to express or withhold their thoughts during conversations, derived from our think-aloud study. Each heuristic contains two example mid-level themes from our codebook.

3.4.3 Findings

Two researchers collaboratively analyzed participants' responses using the affinity diagram method [59]. We held eight 90-minute coding sessions. In each session, we split the transcripts and reviewed the data together to identify meaningful quotes. For each new quote, we proposed potential groupings into existing clusters or created new clusters through discussion. Agreement between the two researchers was required before assigning a quote to a cluster. In cases where consensus could not be reached, the quote was temporarily set aside and revisited during subsequent iterations. Once the initial clusters were established,

we labeled each cluster with a theme. Conflicts in grouping or interpretation were resolved through discussion in the context of the original data. We in total derived 10 high-level themes, 23 mid-level themes, and 68 low-level themes derived from 394 quotes (Figure 3.1). The complete codebook is available in the Supplementary Material.

3.4.3.1 What Thoughts Do People Formulate?

Consistent with dual-processing theory [43], participants reported two types of thoughts. System 1 thinking is fast, automatic, and intuitive, often leading to immediate responses. In contrast, System 2 thinking is slower more deliberate. Participants indicated they use both modes—sometimes responding spontaneously (System 1), while at other times reflecting more deeply before engaging (System 2).

3.4.3.2 Why do people express or withhold a thought?

We summarize eight heuristics to determine whether a participant wants to express or withhold a thought (Figure 3.1), and collectively name them the **intrinsic motivations** for participants to engage in conversations.

Among the most frequently mentioned motivations, **relevance** (77 mentions) emerged as a dominant factor. Participants were more inclined to contribute when topics aligned with their knowledge, interests, or past experiences, resonated with prior long-term memories, or built on their recent thoughts. In contrast, participants often withheld their input when they perceived a disconnect from the ongoing discussion. The role of relevance in conversational engagement aligns with Grice's Cooperative Principle and the maxim of relevance [51]. Similarly, Duncan and Fiske observed that conversational contributions depend on aligning with shared context and ongoing topics [39].

The presence of an **information gap** (33 mentions) also strongly motivated expression. Participants spoke up when they identified missing knowledge, confusion, or the need for

clarification. Addressing these gaps often enriched the conversation with additional details or counterpoints. Conversely, participants withheld their thoughts when they deemed the discussion predictable or unengaging. While our finding is in group conversation settings, this draws parallels with Berlyne's theory of epistemic curiosity, which describes how individuals seek information to resolve uncertainty [11].

The **expected impact** (23 mentions) of a thought further influenced engagement. Participants were more likely to contribute if they anticipated that their input would introduce novel topics, steer the conversation, or enhance its depth. They hesitated when they believed their thought would be redundant or covered later.

Urgency (14 mentions) played a decisive role when participants felt their input was time-sensitive or critical for addressing errors or misunderstandings. Participants expressed thoughts promptly when they perceived such moments as pivotal for the direction of the conversation. While urgency has been discussed in contexts like problem-solving or crisis communication [126], our study identifies its role in everyday group conversational dynamics, particularly in mitigating misunderstandings or addressing immediate errors.

In terms of conversational structure, **coherence** (30 mentions) shaped decisions. This aligns with Sacks et al.'s observation of how speakers organize their contributions to maintain conversational flow [114]. Participants expressed thoughts that logically built upon the previous utterance or extended the topic, while withholding ideas that might disrupt conversational flow. Similarly, **originality** (16 mentions) guided engagement, as participants avoided redundancy by refraining from reiterating points already raised.

Balance (33 mentions) relates to the dynamics of conversation. Participants were mindful of their own contributions relative to others and often sought to maintain inclusivity, encouraging quieter members to speak or refraining themselves to allow others space to participate. This draws parallel to Goffman's theory of face-work [48] and Brown and Levinson's politeness theory [21] about how speakers modulate contributions to preserve group harmony.

Finally, **dynamics** (30 mentions) highlighted the interplay of active participation and silence. Participants were more likely to initiate new topics or fill conversational pauses to sustain momentum. However, they often withheld their thoughts when others were actively speaking or appeared likely to contribute soon, reflecting a sensitivity to conversational flow and timing.

While this study's aim was to identify factors influencing thought expression that can be leveraged in our framework, rather than to exhaustively catalog all possible motivations and create a definitive taxonomy, these themes reveal the multifaceted nature of engagement in multi-party conversations. Rather than simply reacting to the flow of dialogue or previous utterances, participants considered a combination of personal motivations, conversational dynamics, and social considerations when deciding proactive participation.

3.4.3.3 Levels of intrinsic motivation.

We also propose five levels of intrinsic motivation, *i.e.*how strongly and likely one would want to express a particular thought and participate in the conversation.

- **Very Low**: The participant is unlikely to express the thought and participate in the conversation at this moment. They would not express it even if there is a long pause or an invitation to speak.
- **Low**: The participant is somewhat unlikely to express the thought and participate in the conversation at this moment. They would only consider speaking if there is a long silence and no one else seems to be taking the turn.
- **Neutral**: The participant is neutral about expressing the thought and participating in the conversation at this moment. They are fine with either expressing the thought or staying silent and letting others speak.
- **High**: The participant is somewhat likely to express the thought and participate in

the conversation at this moment. They have a strong desire to participate immediately after the current speaker finishes their turn.

- **Very High**: The participant is very likely to express the thought and participate in the conversation at this moment. They will even interrupt others who are speaking to do so.

The five levels of intrinsic motivation serve as output labels for predicting intrinsic motivation in our framework.

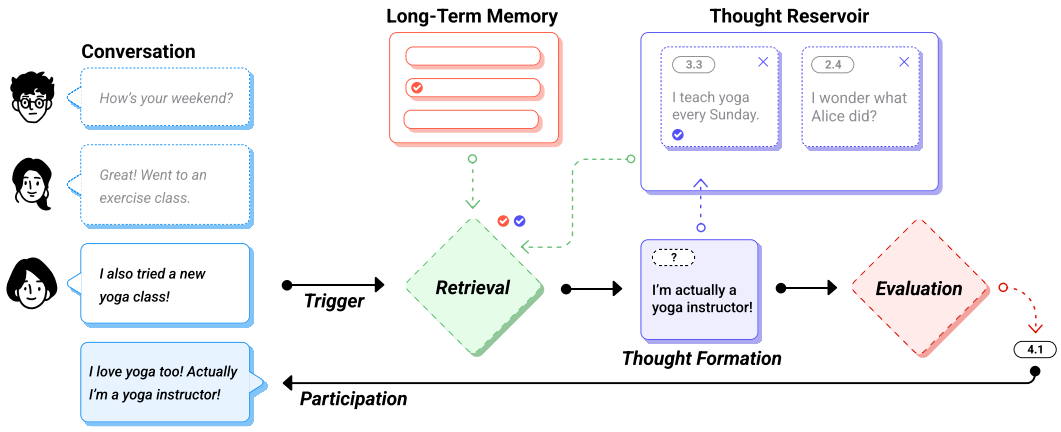


Figure 3.2: The *Inner Thoughts* framework for AI proactive engagement in conversations. A conversational event triggers the retrieval of relevant memories from long-term memory and thought reservoir. New thoughts are then formed based on these activated memories, and added to the thought reservoir. These thoughts are evaluated for AI's intrinsic motivation (score = 4.1 in the figure) to express. AI participates by articulating a thought at a selected moment in the ongoing conversation.

3.5 Inner Thoughts Framework

Motivated by our formative studies, we introduce a computational framework, **Inner Thoughts**, that enables AI proactivity by continuously generating a train of thoughts alongside the on-

going conversation and autonomously deciding when and how to engage.

Our design is inspired by cognitive architectures like SOAR [75] and ACT-R [109], which maintained short- and long-term memories filled with symbolic structures, and operated in perceive-plan-act cycles. These systems dynamically perceived their environment and matched it with pre-defined action procedures. Similarly, our AI retrieves relevant memories, forms thoughts, and evaluates responses in continuous cycles.

The Inner Thoughts framework consists of five components: *Trigger*, *Retrieval*, *Thought Formation*, *Evaluation* and *Participation* (Figure 3.2). In each cycle, a conversational event triggers AI to retrieve relevant memories, form new thoughts, evaluate if there are thoughts that are motivated to be expressed, and then participate by articulating the selected thought at a selected moment in the conversation. The AI will repeat this process as the conversation proceeds.

In our implementation, we chose the hyperparameters described in the following paragraphs empirically to illustrate the core concept of the framework and to demonstrate that it functions effectively within the values chosen. We recognize conducting formal ablation studies an important direction for future work.

3.5.1 Trigger

In human conversations, thoughts often arise in response to specific triggers. Our Inner Thoughts framework mirrors this process by treating conversational events as triggers that initiate AI's internal thought generation. A trigger can take many forms – such as a new utterance, a pause in conversation, a non-verbal cue, or even a keyword embedded within a participant's speech. Any of these events can stimulate the AI to initiate a new thought process and generate a new batch of thoughts.

In the implementation of our system for online text-based conversations, we defined two types of triggers. (1) `on_new_message`: This trigger is activated whenever one of the

participants sends a message. Each incoming message prompts AI to generate a new set of thoughts in response to the latest input. (2) `on_pause`: The second type of trigger occurs when no participant has spoken for a period of time (set to 10 seconds in our system). This allows the AI to generate thoughts during moments of silence, potentially facilitating the interaction by proposing new topics or re-engaging participants. For instance, in our experiment, we observed AI generating thoughts like: "It has been ten seconds and no one has spoken – perhaps I should suggest a new topic?".

3.5.2 Retrieval

Once triggered, AI retrieves information from its memories to use as the *stimuli* to form thoughts. From our think-aloud study, participants mentioned that this could involve long-term memory of related personal experiences, objectives, knowledge, or interest, as well as working memory for details from the ongoing conversation, or even previous thoughts they had. Random memories can also be retrieved to simulate the process of being "creative".

We retrieve relevant memories by computing their *saliency* with respect to the latest utterance. Memories with saliency higher than a threshold (0.3) will be selected. Let x represent a memory item (e.g., an objective, knowledge, or thought) and u represent the latest conversational utterance. The saliency of a memory x is determined by the maximum similarity between the memory and both the raw text of the utterance u and its interpretation u_{interp}. Specifically:

$$\text{Saliency}(x, u) = \max\left(\text{sim}(x, u_{\text{interp}}), \text{sim}(x, u)\right) \cdot w_x \cdot d_x$$

where $\text{sim}(a, b)$ is the cosine similarity between two embeddings, w_x is the weight of the memory x that can be predefined by users and reflects its inherent importance, and d_x is a decay factor that reduces the saliency of older memories. The decay factor d_x is defined as: $d_x = \lambda^{(t-\tau_x)}$, where λ is the decay rate (0.95), t is the current timestep, and τ_x

is the last time or batch when the memory x was accessed. This formulation ensures that more relevant and recently accessed memories have higher saliency, letting the AI focus on pertinent information when generating thoughts during the conversation.

We include the interpretation of the utterance (u_{interp}) alongside the raw text to capture both the surface meaning and the underlying intent or contextual meanings of what was said. The interpretation is generated by prompting an LLM with the instruction: *Interpret what <name> just said in the context of the conversation and what <name> might be thinking. Be as succinct as possible and use a single sentence.*

3.5.3 Thought Formation

Our framework employs a dual-process model [43] of thought formation, based on our think-aloud study findings (section 3.4). This process involves two systems: *system 1* for quick, automatic responses, and *system 2* for deliberate, contextually-rich thinking. Users can configure how many system 1 and/or system 2 thoughts should be generated for each trigger in one batch.

For system 1, we prompt the LLM to form a succinct thought based on the last utterances in the conversation, such as acknowledgments or expressions of interest. For system 2, we prompt the LLM to generate thoughts based on the retrieved stimuli. Below is a short version of the prompt structure (full prompt in Supplementary Material):

> *"You are provided contexts including the conversation history and salient memories of yourself... Form <num> thought(s) that you would most likely to have at this point in the conversation, given the context. Make sure they are diverse, align with these contexts and are less than 15 words."*

We also prompt the LLM to annotate the stimuli (from a previous thought, utterance or long-term memory) for each thought it generates (as shown in Figure 3.5). This provides a traceable link between the AI's memories and thoughts and make the generation process

more grounded.

Similar to what was observed in reasoning and decision-making tasks [151, 99, 141, 120, 152], we empirically found that LLMs can form reasonable and consistent thoughts based on the stimuli and conversational context. For example: *"I should mention the picnic we had last weekend"*, *"I wonder how long Bob's hike was"*, *"Seeing a bear up close must have been intense!"*.

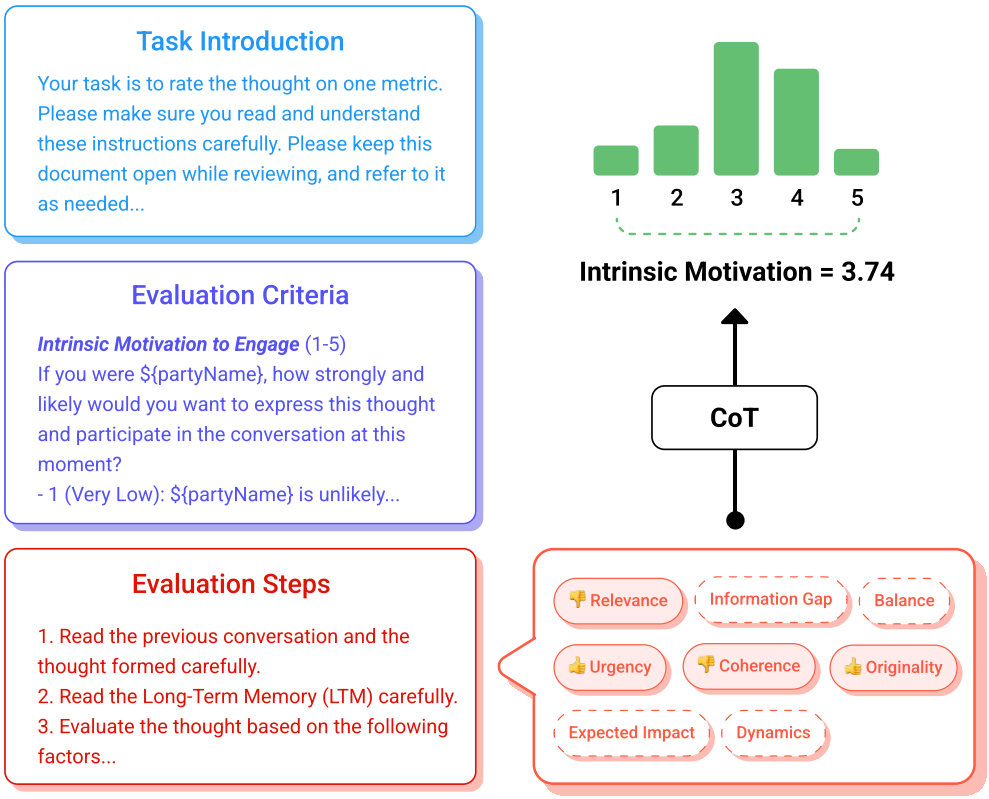


Figure 3.3: Prompt structure for evaluating *intrinsic motivation* of a thought. The evaluator rates the AI's intrinsic motivation to engage using a 1-5 scale based on heuristics like relevance and coherence. A Chain-of-Thoughts (CoT) process evaluates both positive and negative factors, resulting in a weighted score.

3.5.4 Thought Evaluation

Not all generated thoughts will be expressed. In this thought evaluation phase, the AI censors its latest batch of generated thoughts and decides whether or not to express a particular thought.

We use a structured evaluation process. This evaluation is driven by heuristics we developed in section 3.4: *Relevance*, *Information Gap*, *Expected Impact*, *Urgency*, *Coherence*, *Originality*, *Balance* and *Dynamics* (Figure 3.1). Our implementation employs LLMs to evaluate each thought on the set of heuristics and assigns a rating (1-5) to determine the likelihood of the thought being expressed. We also provide definitions of the scores based on the five levels of intrinsic motivation we proposed in section 3.4, from very low to very high. This makes the LLM's prediction grounded and explainable, and can be further used to guide participation strategies.

We developed a pipeline similar to G-Eval [83], a prompt-based evaluation method. Our process involves three key components: (1) a prompt that provides instructions for evaluation and defines the criteria, (2) a structured chain-of-thoughts (CoT) that outlines intermediate steps for evaluation, and (3) a scoring function that computes a final score for each thought based on its probability of being expressed (Figure 3.3).

One unique aspect of our evaluation process that goes beyond the G-Eval is that we instruct the AI to provide both positive and negative motivations for expressing a thought. We empirically found this method to overrate the scores less. This step follows these key instructions:

1. First, reason about why the party may have a strong desire to express the thought and participate in the conversation at this moment. The system selects the top two most relevant factors that may argue for expression (*e.g.*relevance, clarification, or new topic)

2. Then, reason about why the party may have a weak desire to express the thought at

this moment. Again, the system selects the top two most relevant factors that may argue against the expression of the thought (*e.g.*irrelevance, incoherence, or lack of urgency).

3. Based on these considerations, the system assigns a rating on a scale of 1-5 for the motivation to express the thought.

The intrinsic motivation score for each thought is determined using a weighted summation approach inspired by G-Eval [83]. Specifically, we sum the probability scores of the first output token (which is a rating prediction from 1 to 5) of the top five LLM's responses, for every evaluation:

$$\text{score} = \sum_{i=1}^{5} p(s_i) \cdot s_i \cdot d_p$$

Where $p(s_i)$ is the probability of the predicted rating (0-1), calcuated by taking e to the `logprob` (log probability of the output token) value of the LLM response, and s_i represents the predicted rating.

This method allows for more fine-grained, continuous scores compared to traditional integer-based evaluations. The final score is also adjusted by how long the AI has been silent. Our assumption is that in general, the longer a party stays silent, the stronger motivation they will have to participate to maintain their presence. This factor d_p is defined as: $d_p = \lambda^{(t-\tau_p)}$, where λ is the increase rate of motivation score (1.02), t is the current timestep, and τ_p is the last time when party p spoke.

3.5.5 Participation

After evaluating the intrinsic motivation score of its latest batch of thoughts, the AI decides whether to speak by leveraging turn-taking type predictions (*i.e.*turn allocation vs. self-selection), combined with the evaluation of those thoughts. The Inner Thoughts framework allows the AI to exhibit varying degrees of proactive participation through adjustable

proactivity settings. We define three layers of proactivity that control how and when the AI participates in the conversation:

Overt proactivity, which refers to the AI's overall tendency to engage in conversation, similar to how some people naturally participate more actively in discussions, regardless of specific thoughts or ideas. To implement overt proactivity, we adjust the `system1Prob` (System 1 Probability, 0-1) parameter, which controls the probability to select a system 1 thought when no system 2 thought is selected. A higher `system1Prob` increases the chance that the AI will respond in general even when other thoughts are rated to have low motivation.

Covert proactivity, which is the level of motivation required for the AI to express a thought and engage. This is managed through `imThreshold` (1-5), the intrinsic motivation threshold for expressing a thought. A thought may only be selected if it is evaluation score is higher than this threshold.

Tonal proactivity, which shapes how assertive or forward the AI appears in its language. The `proactiveTone` (true or false) controls the AI's style of expression once it has decided to speak. While the core thought-selection process is the same, the proactive tone modulates how assertively the AI conveys its message by restyling the articulated utterance through an LLM.

In addition, Inner Thoughts introduces *interruption*, represented by the `interruptThreshold` (1-5). Interruption occurs when the AI takes a turn despite the turn being allocated to another participant. For example, this might occur when Alice asks Bob, *"How about you, Bob?"* but the AI interjects because it has an urgent thought to express. Interruption is not explicitly outlined in Sacks et al.'s *Simplest Systematics* [114] but is framed here as a mechanism to override the orderly system of turn-taking when necessary. If the intrinsic motivation behind a thought exceeds the `interruptThreshold`, the AI will override standard turn allocation rules to contribute to the conversation. This provides an additional layer of proactivity.

The AI decides whether to speak by leveraging turn-taking type predictions (*i.e.*turn-allocation vs. self-selection) combined with the thought evaluation process. For open turns (self-selection), the AI speaks if its top thought surpasses the intrinsic motivation threshold; otherwise, it may rely on system-1 thoughts or remain silent. For allocated turns, the AI selects its highest-rated thought to speak, and for others' turns, it interrupts only if its motivation exceeds the interrupt threshold. This algorithm is formally described as:

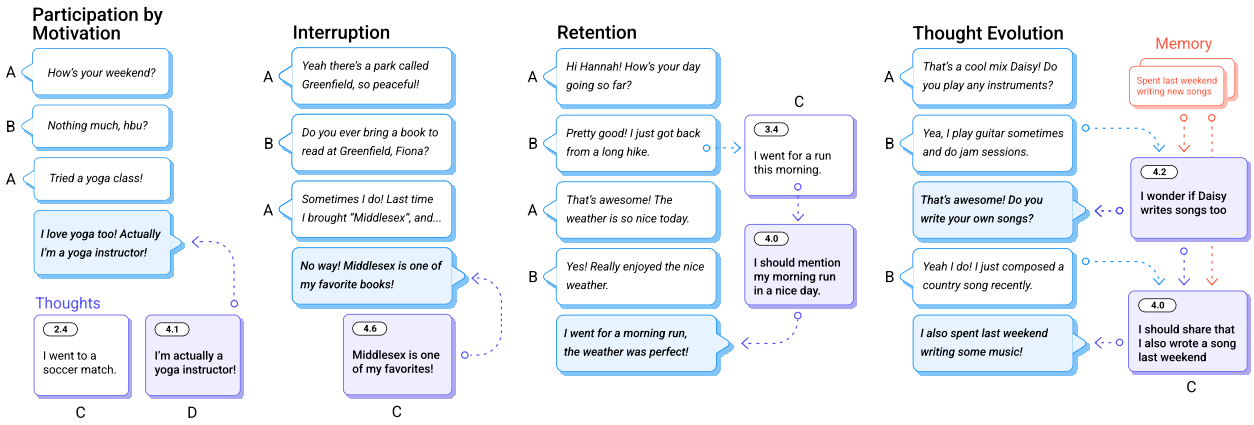


Figure 3.4: Examples selected from simulation logs of AI turn-taking behaviors in the Inner Thoughts framework. The figure illustrates four behaviors: Participation by Motivation, where the AI joins the conversation by sharing relevant personal experience; Interruption, where the AI interjects with a strong contribution during an ongoing discussion; Retention, where the AI holds back a thought until it's contextually relevant; and Thought Evolution, where the AI adapts its responses as the conversation progresses.

3.5.6 Demonstration of AI's Proactive Behavior Enabled by Inner Thoughts

We present several examples selected from simulation logs of AI turn-taking behaviors enabled by the Inner Thoughts framework (Figure 3.4).

3.5.6.1 Participation by Motivation

In the Inner Thoughts framework, AI participation is driven by its intrinsic motivation, as opposed to traditional approaches that rely on conversation history. Previous systems might randomly select participants with minimal interest or knowledge in the topic at hand, potentially stagnating the conversation. In contrast, Inner Thoughts ensures that the AI participates with the strongest motivation — whether due to a relevant persona, curiosity, or the fact that they have not spoken in a while — takes the conversational floor. This dynamic leads to more fluid and engaging topic progression, as participants with something meaningful to contribute are naturally more involved. Over time, this accumulation of motivated contributions may develop conversations that are more coherent, engaging, and reflective of the natural flow of human interaction, as shown in our evaluation in section 3.6.

For instance, as shown in Figure 3.4, the AI demonstrates motivation-based participation when a user mentions trying yoga for the first time. With its knowledge of yoga and background as a yoga instructor, the AI promptly responds: *"I love yoga too! Actually, I'm a yoga instructor!"* The AI's motivation to share relevant personal experience ensures a smooth continuation of the conversation.

3.5.6.2 Interruption

The Inner Thoughts framework enables AI to interrupt a conversation when it has a strong motivation to contribute. Even when participants A and B are discussing a particular topic, participant C (the AI) can step in if it identifies a strong, relatable connection to the conversation. This behavior makes conversations more dynamic and allows the AI to share important insights without needing to wait for a turn. In contrast, methods solely dependent on next-speaker prediction often fail to offer the AI opportunities to engage if the conversation converges around two participants.

As shown in the figure example, while A and B are in the middle of a dialogue, the AI

interrupts with, *"No way! Middlesex is one of my favorite books!"* This interruption enriches the conversation by fostering more spontaneous interactions.

3.5.6.3 Retention

In addition to its ability to interrupt, the Inner Thoughts framework also allows the AI to retain thoughts for future use, waiting for an appropriate moment to express them. This feature enables the AI to revisit previously generated thoughts that may have been irrelevant at the time but later become pertinent as the conversation progresses.

For example, in the figure, the AI initially holds back a thought about going for a run earlier that day because it was not particularly relevant while other participants were discussing a different topic. However, once the conversation shifts to the weather and outdoor activities, the AI sees an opportunity to contribute: *"I should mention my morning run in the nice weather."*

3.5.6.4 Thought Evolution:

The Inner Thoughts framework allows for the development and evolution of thoughts over time. Unlike traditional systems that generate responses based on a fixed persona [155, 150], Inner Thoughts enables the AI to develop and adapt its thoughts as the conversation unfolds, incorporating multiple stimuli along the way.

For instance, as shown in the figure, the AI initially recalls a memory of writing songs last weekend. As the conversation shifts toward music and instruments, this memory evolves into the thought: *"I wonder if Daisy writes songs too."* It expresses the thought by asking Daisy the that question. With a positive answer from Daisy, the thought further evolves into: *"I should share that I also wrote a song last weekend."* This continuous evolution allows the AI to stay relevant and responsive as new topics emerge, as well as compound and develop new thoughts.

3.6 Simulative Evaluation

We conducted a technical evaluation via multi-agent simulations to compare different strategies in enabling proactive AI engagement in multi-party conversations. We chose a simulative approach to overcome weaknesses of only relying on conventional user studies with human subjects. First, the difficulty to scale due to the time cost of coordinating human participation. Second, we found through our pilot studies that human participants may struggle to perceive the timing of AI engagement in social conversations, focusing more on the style and content of responses. In addition, since many forms of engagement may seem reasonable in social conversations, participants often do not have clear criteria to assess AI's engagement behavior.

Taking a non-conventional approach, our intuition is that simulating conversations at-scale amongst multiple AIs using the same engagement strategy allow us to accumulate and magnify the effects of both correct and incorrect turn-taking decisions. In particular, poor decisions about when to engage can compound and lead to noticeable degradation in conversation quality, making evaluation more straightforward. This method also offers scalability, as crowdworkers can assess conversation quality without the need for real human interactions with AI.

In this section, we compare the performance of our proposed Inner Thoughts framework with the conventional next-speaker prediction baseline in multi-agent simulations.

3.6.1 Apparatus: the Inner Thoughts Playground

We built an Inner Thoughts playground (Figure 3.5) that allows us to simulate conversations between AI and/or human participants. This playground is deployed at `https://liubruce.me/inner_thoughts/`. On the playground, users can easily add multiple AI and human participants, customize their proactivity settings, control the number of thoughts formulated per batch, run automated simulations of human-AI group conversation, and save log data of

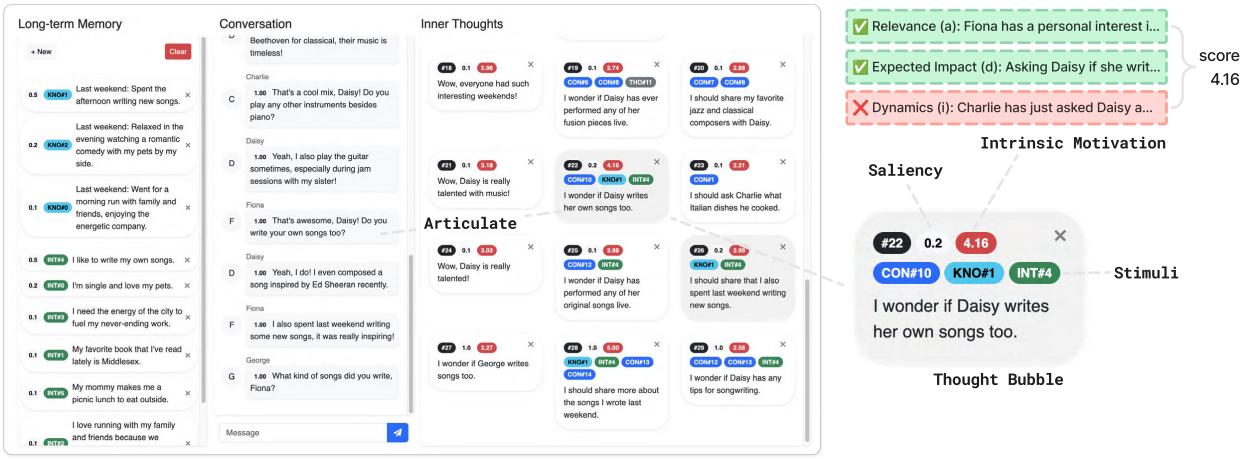


Figure 3.5: The Inner Thoughts playground web app interface. Multiple AI and humans can be added to simulate a group conversation. Users can also view and edit each of the participants' long-term memory and thoughts.

conversations.

The main interface is divided into three panes: On the left is the long-term memory pane. Users can customize each AI participants's long-term memory by adding or deleting specific entries. In the middle is the conversation pane, where users can watch the simulated conversation, or participate in the conversation by sending a message using the dialog box on the bottom.

On the right is the inner thoughts pane. Users can view visualization of thought bubbles generated on-the-fly for the selected participant as the conversation proceeds. Each thought bubble contains a numeric ID (badge colored black), the saliency score (white), intrinsic motivation score (red), and a list of stimuli that AI used to formulate this thought. Below the badges shows a description of the thought. Thoughts that are expressed by the AI will be highlighted. In addition, users can click on a thought bubble to manually force the AI the express the thought in the conversation; and can right click on the thought bubble to view its reasoning for the intrinsic motivation rating. Users can also delete certain thoughts from the reservoir by clicking on the top-right delete button.

As a user clicks on a participant in the settings page, the content in the long-term memory and inner thoughts pane will be synced to that participant. They are automatically updated as the conversation goes on. Users can view a train of thoughts of each participant developing in parallel to the conversation.

3.6.2 Conditions

We compared two multi-party engagement strategies (Table 3.2) (1) Next-Speaker Prediction: In this condition, AI participants engaged based on predictions of who the next speaker would be. Their responses were generated based on predefined personas, following experimental setup in [155]. (2) Intrinsic Motivation (our approach): AI participants engaged based on their intrinsic motivation to contribute, driven by generated thoughts during the conversation.

We used the fine-tuned GPT-3.5 model we evaluated in section 3.3 in condition 1 to predict the next speaker, and prompt the model to generate responses based on its persona if selected by the prediction (full prompt in Supplementary Material). We used the framework described in section 3.5 for condition 2. Simulations were run on the Inner Thoughts playground web app.

3.6.3 Agent Personas and Conversation Generation

In this paper, we choose casual conversation scenarios as the primary focus of our evaluation due to their unique challenges in handling turn-takings. Casual conversations, unlike task-oriented interactions, lack clear objectives, making turn-taking and proactive engagement particularly difficult to model and has been underexplored in prior research [81, 47]. Future research could also explore how this framework applies to task-oriented conversations such as brainstorming, to further validate its adaptability.

To simulate the conversation, we first created eight AI participants, each assigned a

detailed persona consisting of objectives, knowledge, and interests. These personas were initially generated from a seed randomly selected from the PersonaChat [155] dataset. For example, one seed might include: "I like listening to all genres of music except country," "I would travel the world if I could," "I enjoy reading books," "I like spending time with friends and family," and "I'm not a fan of hot weather."

To further enrich these personas and encourage interaction between different AI participants, we randomly sampled two additional persona descriptions from other participants for each AI. This approach introduced overlapping interests, making the AI participants more likely to engage in relatable conversations and increasing the chances of contributions during discussions.

We simulated 100 text-based conversations (50 for each condition), each involving four randomly selected AI participants. Each conversation consisted of 15 turns. We also incorporated 10 icebreaker prompts, randomly selected from the PersonaChat dataset. Examples of these prompts include: "What did you do last weekend?", "What is your favorite thing to do?", and "Hey!". A randomly selected participant was chosen to initiate the conversation for each simulation, with a randomly selected icebreaker sentence. For all AI participants, the following proactivity settings were applied: Overt proactivity = 3.95, Covert proactivity = 0.1 and Tonal proactivity = False. One system 1 thought and two system 2 thoughts are generated in each batch.

3.6.4 Hypothesis

With the Inner Thoughts approach, the AI participant with the highest intrinsic motivation is more likely to take the floor of the conversation. Such participants tend to have more to contribute and are better able to develop the topic at hand. As discussed in subsection 3.5.6, intrinsic motivation fosters more meaningful and contextually appropriate contributions, in contrast to reactive engagement that simply predicts the next speaker. We anticipate that this effect will aggregate and lead to conversations that are more engaging, coherent and

closely resemble the natural flow of human conversations.

3.6.5 Human Evaluation of Simulated Conversations

We evaluated 100 simulated conversations with 10 human evaluators (4 female, 6 male, age Avg.: 26.3, SD: 4.35). Each evaluator reviewed five pairs of conversations, with one from each condition (10 total), viewed in a randomized order within each pair. Participants were informed that all conversations were AI-generated. Instead of displaying static conversation histories, we presented an animated version of the conversations to simulate a live chat experience. The length and speed of the conversations were rendered to match human's average typing speed. This decision was based on findings from our pilot studies, where we observed that participants tended to skim through static conversation transcripts.

After watching each conversation, they were asked to rate their agreements with seven statements (Table 3.3) related to the conversation's quality on a 1-7 Likert scale, from strongly disagree to strongly agree, adapted from [16, 140, 9]. We also asked participants to identify specific points where turn-takings felt unnatural. After completing all conversations, participants were asked to select the conversation that feels more natural and human-like in each pair and provide a brief explanation with examples.

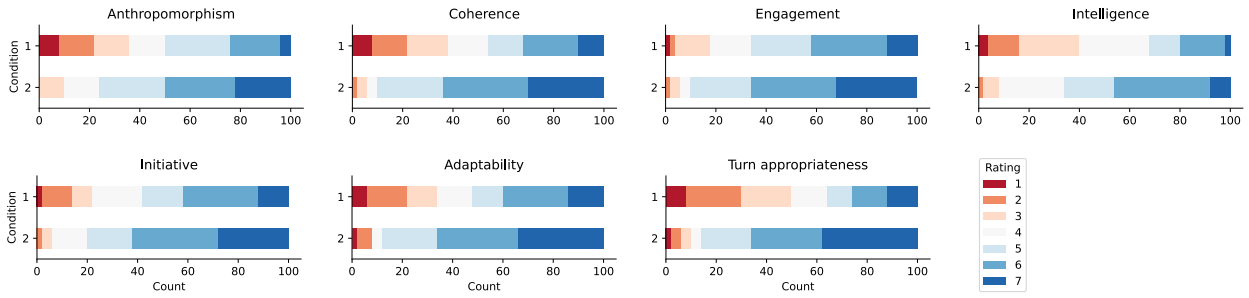


Figure 3.6: Stacked bar plots of participants' ratings on the metrics used in our technical evaluation to measure the quality of AI simulated conversations. Each statement is rated on a Likert-scale from 1 – Strongly Disagree to 7 – Strongly Agree.

3.6.6 Findings

We conducted Mann-Whitney U tests to compare the Baseline (condition 1) and Inner Thoughts (condition 2) strategies across the 7 dimensions listed in Table 3.3. Overall, the results showed significant improvements in the Inner Thoughts condition across all metrics (Figure 3.6).

Notably, the strongest effects were observed in turn appropriateness (U = 577.0, $p = 2.4 \times 10^{-6}$) and coherence (U = 636.0, $p = 1.6 \times 10^{-5}$), indicating that AI participants using intrinsic motivation contributed more appropriately and maintained better conversational flow compared to the Baseline. Anthropomorphism (U = 726.5, $p = 2.4 \times 10^{-4}$) and intelligence (U = 688.5, $p = 7.3 \times 10^{-5}$) were also significantly higher for Inner Thoughts, reflecting that AI participants were perceived as more human-like and thoughtful.

The Inner Thoughts condition also led to significantly higher engagement (U = 813.5, $p = 1.9 \times 10^{-3}$) and initiative (U = 862.5, $p = 6.0 \times 10^{-3}$), suggesting that AI participants were more proactive and engaging in conversation. Adaptability was rated significantly better in the Inner Thoughts condition as well (U = 765.0, $p = 6.3 \times 10^{-3}$), showing that AI participants adapted more effectively to changes in the conversation. Finally, 82% of the times participants preferred the Inner Thoughts conversations, indicating a clear overall preference for this approach in multi-party settings.

To complement these quantitative findings, we analyzed participants' feedback to understand the nuances behind their ratings. We used a thematic analysis approach [18] to analyze qualitative data from participant comments. Two researchers first independently identified recurring themes and patterns in the data, followed by two 60-minute meetings to refine and organize these themes. Disagreements were resolved through discussion.

Enhanced Coherence and Engagement Participants noted that conversations in Condition 2 (Inner Thoughts) felt more coherent and engaging. They appreciated how AI agents

built upon each other's responses, creating a more dynamic and interactive dialogue.

> *"In the second conversation, every participant adjusted their answers based on others' responses. It started general and then narrowed down, making it more engaging."* (P01)

Natural Turn-Taking Condition 2 was praised for its natural turn-taking, with AI agents contributing at appropriate moments and responding directly to others.

> *"It felt like a real group chat where participants are listening to each other and interested in each other's topics. Their interaction is closer."* (P06)

> *"There was a flow in the conversation—not mechanical. If someone mentioned something, others would continue on that topic, echoing and adding more information."* (P08)

Responsiveness to Context Participants observed that AI agents in Condition 2 were more responsive to the conversational context, leading to more meaningful interactions.

> *"... they can combine with their own experiences, making the conversation feel more connected."* (P04)

> *"Conversations had natural transitions. You feel like they are responding first and then sharing something about themselves."* (P08)

Limitations of the Baseline Strategy Conversely, the baseline condition was criticized for its mechanical responses and lack of coherence. Participants felt that AI agents often talked past each other without meaningful engagement.

"In the first conversation, everyone was talking over each other. They answered the same question with the same format, which felt unnatural." (P07)

"Everyone says 'hey' but there's no continuity. They don't respond to each other and just bring up unrelated topics." (P09)

Missed Opportunities for Interaction The baseline AI agents frequently failed to respond to prompts or engage with others' statements, leading to disjointed conversations.

"Someone mentioned walking their dog on the beach, and no one responded at all." (P03)

"They ignored questions, and some people went back to very previous messages, making the conversation one-directional." (P10)

3.7 User Evaluation

In addition to simulation experiments, we conducted a user study to understand: (1) How do people perceive proactive conversational AI enabled by the Inner Thoughts framework during actual interactions? (2) How do different levels of AI proactivity affect these perceptions?

3.7.1 Apparatus: *Swimmy* Slackbot

We built a Slackbot (Figure 3.7) named *Swimmy* using Slack API[3] for the study. We implemented the Slackbot using a queue-based approach to handle asynchronous message processing. When a new message is received, it is added to a `triggerQueue`, and processes each message sequentially. The bot generates inner thoughts for each of the AI participants by updating saliency, forming thoughts, and evaluating them, as described in section 3.5.

[3]https://api.slack.com/

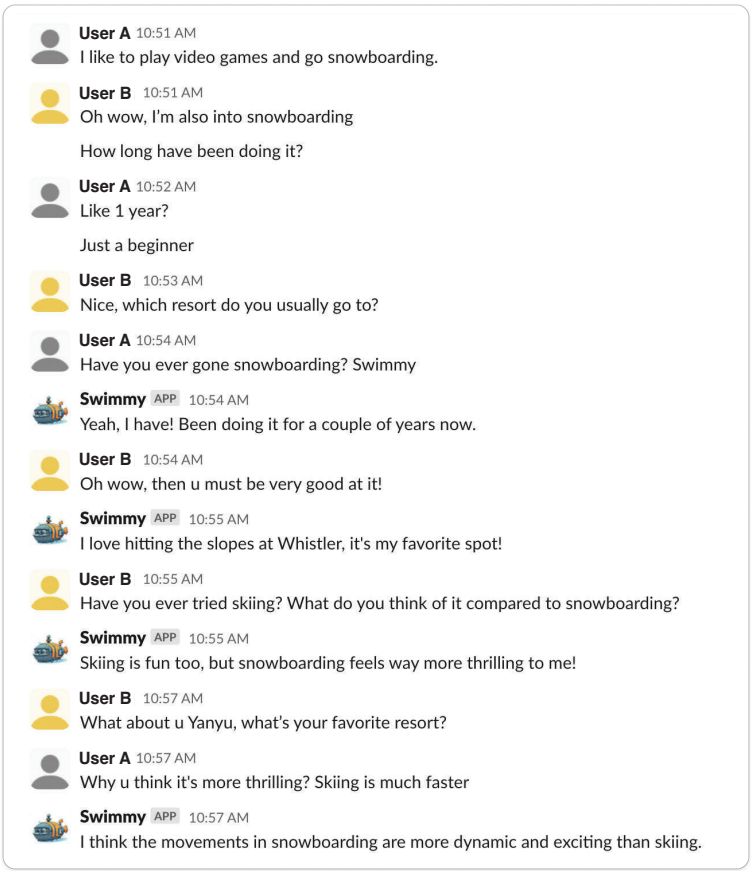


Figure 3.7: An example conversation between two human participants and *Swimmy*, a Slackbot developed based on the Inner Thoughts framework, on Slack.

The AI then decides whether to respond based on turn-taking predictions, intrinsic motivation thresholds, as well as the status of the process queue. Specifically, if `triggerQueue` is not empty, the AI refrains from speaking to avoid interrupting ongoing message processing. Users can also customize configurations on the *Home* page of the Slackbot.

3.7.2 Study Design

The study involved six pairs of human participants, each pair interacting with an AI agent on Slack (3-party conversation). We recruited 12 participants from our institution, with 8 of them reporting familiarity with conversational agents and 9 indicating familiarity with large language models (scoring above 4 on a 1-7 scale). Each participant was compensated $20 for their one-hour participation.

Participants experienced three 10-minute conversations, where we designed three AIs with different conversational styles:

1. *Non-stop chatter*: This AI participated continuously, even when it had little input to add to the conversation. It had a high probability of selecting thoughts through System 1 processes (`system1Prob` = 0.7).

2. *Active contributor*: This AI participated actively to share its thoughts whenever relevant but not overwhelming the conversation. With a moderate System 1 probability (`system1Prob` = 0.2), it occasionally selected less deliberate thoughts. Its low intrinsic motivation threshold (`imThreshold` = 3.59) allowed it to contribute actively whenever something comes up in its "mind".

3. *Selective participant*: This AI only contributed when it had a strong interest in the topic, staying quiet during other parts of the conversation. With no reliance on System 1 processes (`system1Prob` = 0) and a higher motivation threshold (`imThreshold` = 4.09), this AI required significant justification before expressing a thought.

Each persona was randomly assigned to one of the three conversational conditions in each session, with order counter-balanced. Before the study, participants were informed that the AI might exhibit various personalities across the conversations but were not told in advance about the specific conditions.

We prompted participants to have a casual chat on three social topics: hobbies and interests, travel experiences, and weekend activities. After each conversation, participants rated their experience on ten metrics adapted from [16, 140, 9]. Compared to technical evaluation, we added metrics related to how people perceive their interaction with the chatbot, like *Likeability*, *Social Presence*, *Perceived Listener*, *Contribution*, etc.

At the end of the study, participants were asked to match the chatbots to the behavior styles listed above, and engaged in a semi-structured interview about their experience. They were also asked to select their favorite AI out of three conditions.

3.7.3 Findings

3.7.3.1 Interacting with Proactive AI

Overall, participants had a positive perception of the proactive conversational AI enabled by the Inner Thoughts framework. Across all three conditions, the AI was rated to have median scores of 5 for anthropomorphism, initiative, engagement, listening ability, contribution, and extroversion. Notably, likeability and response timing also leaned toward positive ratings, with a median of 6, indicating that many participants found the AI likable and can engage at appropriate timings.

3.7.3.2 Perception of AI Proactivity

To evaluate whether the designed proactivity styles were both perceptible and distinguishable to users, we asked participants to identify which AI exhibited each conversational style after their interactions. The accuracy of their guesses varied, with the Non-stop Chatter being

correctly identified 69.23% of the time, making it the most easily recognized. In contrast, the Selective Participant and Active Contributor were correctly identified 54.55% and 50% of the time, respectively. All AIs were identified above the baseline accuracy of 33.3%, indicating that participants were able to distinguish the AIs at a level higher than chance.

Participants also reported different perceptions of AI with different proactivity levels. For instance, for Selective Participant, participants noted that the AI was often too passive. It would only speak when prompted and failed to contribute actively, with P04 describing the AI as paradoxical: *"It responded enthusiastically when directly addressed, but was otherwise disengaged."* This lack of proactivity led some to feel the AI was overly focused on information retrieval, rather than maintaining a more human-like balance of social interaction and contribution P01.

In contrast, Condition 2 (Active Contributor) was appreciated for being more balanced. Participants praised its ability to engage naturally and at appropriate moments. P02 noted that it tended to initiate new topics when the conversation was drying up. P03 found that the AI was more proactive in encouraging others to share, even calling people by name to invite participation. However, there were still occasional lapses, as others observed that the AI sometimes failed to fully grasp the context of the dialogue, responding in ways that felt slightly off (P07).

For Condition 1 (Non-stop Chatter), while it was often described as overly talkative, participants differed on whether this was seen positively or negatively. P01 felt it mirrored conversations with a group of friends who are excited to chat, but noted that its excessive contributions disrupted the conversational flow, as it tended to speak over others or introduce irrelevant topics. P10, P12 found this persona too overwhelming, describing it as failing to respect the natural pauses of a conversation.

Interestingly, while participants could differentiate between these personas, their ratings for each were not significantly different overall. However, we did observe two statistically significant differences. The Non-stop Chatter received the highest ratings for perceived social

presence (Median = 6), reflecting its continuous participation. In contrast, the Selective Participant was rated significantly lower than the Non-stop Chatter in both perceived social presence (Median = 5.5, $p < 0.05$) and extroversion (Median = 4.5, $p < 0.05$), which is consistent with its more reserved, less engaging behavior.

3.7.3.3 Preferences Over AI Proactivity

Clear preferences for AI proactivity emerged in the study. Condition 2 (Active Contributor) was the most favored, with 6 participants selecting it as the best. Participants appreciated its balanced approach, noting that it contributed actively without overwhelming the conversation. One participant P02 emphasized that this AI was more respectful of conversational flow, waiting for appropriate moments to introduce new topics, making interactions feel more natural.

Condition 1 (Non-stop Chatter) received mixed feedback, with 4 participants rating it as their favorite, but many also criticizing its excessive, often irrelevant contributions. P07 observed that the AI's tendency to introduce long and unnecessary questions disrupted the natural rhythm of human interaction, while another P02 mentioned that the AI reminded them of a language exchange partner who was overly eager to contribute but lacked an understanding of the social context.

Condition 3 (Selective Participant) was the least preferred. Only 2 participants selected it as the best, while 7 rated it as the worst. Participants generally felt that this AI was too passive, contributing little to the conversation unless directly asked. Some P06, P11 noted that it seemed uninterested or unmotivated to engage with the topic at hand, leading to more disjointed conversations.

3.8 Discussion and Future Work

3.8.1 Proactive Agents via Intrinsic Motivation vs. Extrinsic Cues

In this paper, we introduced the Inner Thoughts framework, which emphasizes the role of intrinsic motivation in enabling proactive conversational agents. Unlike traditional approaches that rely primarily on external cues such as turn-taking predictions, our framework explores how AI can leverage internally generated thoughts to determine participation in conversations. While we highlighted the inherent limitations of next-speaker prediction strategies and demonstrated the advantages of integrating intrinsic motivation, it is crucial to note that these methods are *not mutually exclusive*. Instead, they are complementary components that, when combined, could create more robust results. We advocate for the development of holistic systems that integrate internal processes (such as thought evaluation and intrinsic motivation) with external strategies (like multimodal cues and turn-allocation mechanisms [142, 34, 42, 101, 102, 65, 89, 13, 14, 15]). Future research should focus on understanding how these internal and external elements interact and how their synergy can enhance both the functionality and user experience of conversational agents.

3.8.2 Applying Inner Thoughts Beyond Casual Conversations

While our study primarily explored the Inner Thoughts framework in casual conversational settings, its potential extends beyond this domain. The framework's inherent adaptability allows it to be adapted into task-oriented scenarios such as brainstorming, coordination, and negotiation. By customizing the criteria used for thought evaluation and aligning them with the goals of a given scenario, the Inner Thoughts framework can enable goal-oriented proactivity. For instance, in a negotiation setting, the AI could evaluate its intrinsic motivation to contribute based on criteria such as the strategic value of its input or its alignment with pre-defined negotiation strategies. Similarly, in coordination tasks, the framework could prioritize thoughts that promote alignment among team members or clarify ambiguities. Fu-

ture work could explore refining the framework's adaptability by integrating domain-specific heuristics and dynamically learning from user feedback, further enhancing its applicability to real-world task-oriented interactions.

3.8.3 Proactive Conversational Agents Beyond Text and Computational Efficiency

Another exciting avenue for the Inner Thoughts framework lies in expanding its implementation beyond text-based interactions. Extending the framework to support multimodal communication like audio and face-to-face interactions introduces both opportunities and challenges. Real-time, multimodal systems must contend with lower latency requirements and more complex turn-taking mechanisms that incorporate additional cues such as intonation, gestures, and facial expressions. To achieve lower latency in these multimodal systems, simplifying the thought evaluation process by focusing on core metrics like relevance and coherence could be effective. Alternatively, leveraging advanced techniques like training lightweight LoRA (Low-Rank Adaptation) [61] models might strike a balance between computational efficiency and performance, enabling the framework to operate in real-time. Future work should evaluate these approaches to identify optimal strategies for scaling Inner Thoughts to multimodal and low-latency environments, unlocking its full potential for human-like interaction.

3.8.4 Technical Limitations of the Inner Thoughts Framework.

A key limitation of the Inner Thoughts framework is thought formation. The AI sometimes generates irrelevant or contradictory thoughts, which we tried to address by using stricter prompts. However, this could make the thoughts generated repetitive. Future work should explore more advanced methods, such as incorporating knowledge graphs to improve the thought generation process.

Another issue is setting proactivity thresholds. Currently, these thresholds are adjusted through trial and error, leading to inconsistent interactions. Future work could explore a data-driven approach, for example using reinforcement learning to dynamically learn thresholds based on user feedback.

The thought evaluation process also needs refinement. While effective in most cases, the AI occasionally misses opportunities to engage due to underestimating motivation scores, or interrupts too abruptly when motivated to speak. Future iterations of the framework could benefit from more robust evaluation mechanisms that better balance engagement opportunities with conversational appropriateness, potentially integrating adaptive learning techniques to fine-tune these processes over time.

3.8.5 Automatic Evaluation for AI Proactivity

Evaluating the quality of AI engagement in multi-party, non-task-oriented conversations presents unique challenges, primarily due to the absence of clear objectives or predefined outcomes. This inherent ambiguity complicates the development of standardized success metrics, making traditional evaluation methods less effective. HCI research has relied heavily on user studies to assess conversational AI performance [13, 14, 15]. While these studies offer valuable qualitative insights into user experiences, they face limitations in scalability and often lack the precision required for setting granular, reproducible benchmarks. In our work, we tried to address these challenges by conducting human evaluations of simulated multi-party conversations. However, this approach is still relatively resource-intensive and challenging to replicate at scale. This underscores the need for future research to establish robust, cost-effective benchmarks and develop automatic metrics in evaluating AI proactivity. Recent work in NLP, such as leveraging LLMs for self-assessment [83], could inspire comprehensive evaluation frameworks. By establishing such metrics, we can better quantify AI performance, reduce reliance on human evaluations, and enable the systematic development of proactive conversational AI systems.

3.8.6 Other Applications of Inner Thoughts

The Inner Thoughts framework holds exciting potential for a variety of applications. One intriguing use case is in brainstorming sessions, where AI could ambiently generate and suggest ideas, mirroring the thought processes of human participants and offering spontaneous contributions. Additionally we could allow for further customization – such as tuning the AI's thoughts to be more creative, whimsical, or even deliberately childlike. This ability to modify the AI's internal reasoning opens possibilities for specialized applications.

Moreover, the framework introduces opportunities for simulating complex AI behaviors, such as ethical dilemmas or conflicting thoughts. For instance, an AI could be designed to decide whether to lie or reveal a difficult truth, simulating moral decision-making processes. In this way, the AI could offer nuanced interactions that reflect more sophisticated social and ethical considerations, providing a deeper simulation of human-like cognitive behavior.

3.8.7 Interacting with the Thoughts of LLMs

With the Inner Thoughts framework and recent LLMs like Chain-of-Thought (CoT) prompting [141] and OpenAI o1 [98] that leverage internal reasoning processes, a new design question emerges: how could humans interact with the inner thoughts of LLMs? The concept of interaction paradigms for LLMs' thoughts opens up a number of possibilities and challenges. Instead of simply providing outputs, LLMs could surface their intermediate reasoning, enabling users to gain insights into the model's decision-making process. For instance, should these "thoughts" be visible in real time, offering users a glimpse into the system's reasoning trajectory? Further, how might we design interfaces that allow users to question, refine, or even contribute to the AI's inner thought process? These considerations not only affect usability but also trust, as understanding the AI's rationale could make its behavior more transparent and predictable. Addressing these questions will be crucial in defining the next generation of human-AI collaboration paradigms.

3.9 Conclusion

In this paper, we presented the Inner Thoughts framework, a novel approach to proactive AI in multi-party conversations. Unlike traditional systems that rely on turn-taking predictions, our framework enables AI to generate and evaluate its own internal thoughts continuously, deciding when and how to engage based on intrinsic motivation. Our evaluations demonstrated that AI guided by Inner Thoughts offers more natural, engaging, human-like turn-taking behaviors compared to the next-speaker prediction baseline. Our implementation, showcased in a web app and a chatbot, highlights the potential of this framework for future applications in multi-party conversational systems. Our work contributes a novel perspective on proactive AI in conversational settings, highlighting the importance of internal thought processes and intrinsic motivation.

Algorithm 2: *

Input: Stream of trigger events E, Thought reservoir T, Turn-taking type prediction.

Output: Participation action t^* for each trigger event.

while *there is a new trigger event* $e \in E$ **do**

`// Step 1: Turn-Taking Type Prediction`

Predict the turn-taking type for the current event e: Open to anyone, or allocated to a party;

`// Step 2: Process According to Turn-Taking Type`

if *open to anyone* **then**

if $\exists t \in T$ *such that* $score(t) \geq imThreshold$ **then**

Select the highest-rated thought: $t^* = \arg\max_{t \in T} score(t)$;

else if $\nexists t \in T$ *such that* $score(t) \geq imThreshold$ **then**

Select t^* from the system-1 thoughts in T with probability $system1Prob$;

if *turn allocated to AI* **then**

Select the highest-rated thought: $t^* = \arg\max_{t \in T} score(t)$;

if *turn allocated to others* **then**

if $\exists t \in T$ *such that* $score(t) \geq interruptThreshold$ **then**

Select the highest-rated thought: $t^* = \arg\max_{t \in T} score(t)$;

else if $\nexists t \in T$ *such that* $score(t) \geq interruptThreshold$ **then**

The AI remains silent;

`// Step 3: Finalize Participation`

if t^* *is not null* **then**

Participate with t^*;

else

Take no action;

Proactive AI via Inner Thoughts: iteratively processes trigger events by predicting turn-taking types, evaluating thoughts in a reservoir, and deciding whether to participate or remain silent.

Condition	When to Participate	What to Say
1	Next speaker prediction	Based on persona
2 (ours)	Intrinsic Motivation	Based on thoughts

Table 3.2: Study conditions for technical evaluation and user study. We compare Inner Thoughts with the baseline approach of deciding when to participate by next speaker prediction, and then generate response based on AI's persona

Metric	Statement
Anthropomorphism	I felt the conversation is natural and human-like.
Conversation Coherence	I felt that the dialogue maintains a coherent topic progression.
Perceived Engagement	I could feel that the AI participants are engaging well in the conversation.
Perceived Intelligence	I felt that the AI participants provided intelligent and insightful contributions to the conversation.
Turn Appropriateness	Turn-takings in the conversation is contextually and logically appropriate
Initiative	I felt the AI participants are able to take the initiative in conversations.
Adaptability	I felt that the AI participants appeared to adapt well to the changing dynamics of the conversation.

Table 3.3: Metrics used in our technical evaluation to measure the quality of AI simulated conversations. Each statement is rated on a Likert-scale from 1 – Strongly Disagree to 7 – Strongly Agree.

CHAPTER 4

From Transient to Persistent: A Text-Native Interface for Generative Video Authoring

Chapters 2 and 3 focus on communication structure (parallel channels) and initiative (proactive engagement). Chapter 4 addresses the third constraint of turn-taking: *transience*. Chat-based interaction tends to treat collaboration as a sequence of ephemeral exchanges, where plans, decisions, and artifacts exist primarily as text scrollback. For complex creative work—such as authoring a multi-shot video—this transience becomes a practical limitation: creators need a shared representation that accumulates structure, supports revision, and makes both human and AI actions legible over time.

Doki explores a persistent alternative: instead of producing a locked output (*e.g.*a rendered video or a fixed timeline), the system centers a *document representation* as the shared working space between human and AI, designed to remain revisable and easy for users to intervene in. To make this persistent space operational, Doki adopts a structured mapping from writing to video production—*document as video, paragraph as sequence, sentence as shot*—so that authors can build narrative and production structure directly in a single evolving artifact.

4.1 Introduction

Writing gives form to our imagination and stories. Children are taught the basics of narrative structure - beginning, middle and end – as early as first grade. Over time, writing text turns

into a core means for communication and self-expression, as we compose emails, articles, or text messages. With the rise of video platforms like YouTube and TikTok, video has emerged as a new powerful alternative for communication, expression, and storytelling. And recent advances in generative AI introduce new possibilities for content creation. As a result, more and more people are authoring video.

Unfortunately, authoring video remains challenging, because it relies on a very different set of tools than the ones we learn in school. Traditional video non-linear editors emphasize rich interfaces with many capabilities excelling at fine-grained editing tasks at the cost of usability and accessibility. Modern editors, such as Descript [36] reduce friction by supporting text-based editing with transcript text, yet such tools only use text for editing speech tracks and adopt the experience of traditional tools with multiple panes and multi-track timelines for authoring visuals.

Generative AI introduces a paradigm shift in creative tools. Many call generative models "the new camera", because they let people produce high-quality content with a simple text prompt. In this new era, **text is again a central medium for creative expression**. While most generative video tools [50, 20, 113, 103, 70] focus solely on individual clips and still rely on traditional workflows and timelines for narrative storytelling, we ask: if audiovisual content can be generated directly from text, **can video authoring become as natural as editing a document?**

We present *Doki*, a text-native interface for authoring generative videos[1]. In Doki, writing is the primary interaction. Within a single document, users define assets, structure scenes, generate shots, refine edits, and add audio. The document serves both as narrative and as an executable script for video production. Our design pursues four goals:

(1) *Make text a central medium for authoring.* Text is both natural for people and native

[1]We use the term *generative videos* to denote content produced with generative AI methods. Unlike "AI-generated videos," which can imply AI-led authorship, *generative videos* is neutral about agency and ownership.

to AI. It allows humans to freely edit and quickly review, while enabling AI to generate and suggest within the same medium. Its freeform nature also supports flexible workflows that move fluidly between ideation, generation, and revision.

(2) *Consolidate authoring into a single representation.* Current workflows require creators to juggle multiple forms and views of the same video, increasing cognitive load [25]. In Doki, scripts, prompts, visuals, audio, and timelines co-exist in a structured text representation. While we do not aim for a full replacement for professional toolchains, Doki supports many end-to-end creative tasks.

(3) *Preserve consistency via parameterization.* As projects scale, maintaining coherence becomes difficult. Doki introduces parametrized definitions to keep characters, styles, and other elements consistent across narratives, and makes it easier to construct complex scenes and develop longer, more cohesive stories.

(4) *Keep interaction simple.* Professional tools are powerful but can be overwhelming with complex UI. Doki offers a minimalist design with a lightweight slash menu for creating assets and shots, and inline shot previews for reviewing and editing directly in the document.

We evaluated Doki in a week-long diary study with 10 participants, spanning filmmakers to first-time creators. Participants submitted 46 videos and rated Doki System Usability Scale (SUS) 81.2 – "Excellent" – describing faster idea-to-content flow, easier coherence via parameterization, and clearer comprehension of story structure in the document view. We also explored tradeoffs from Doki's text-native interface.

In summary, we contribute:

- **A structured, parametrized text-based representation for videos**. We propose a three-layer, hierarchical *"document as video, paragraph as sequence, sentence as shot"* structure, and introduces a definition system (@mentions and #hashtags) with propagation, scoping, and context handling. This representation is independent of any particular interface.

- **Doki, a minimalistic, text-native interface** that instantiates this representation. Users can create videos end-to-end, from ideation to export, within one single document. Doki employs a minimalistic user interface design with a lightweight slash menu, inline previews and AI agents.
- **An in-the-wild evaluation of Doki**. We report findings from a week-long diary study that examined how participants used Doki. We identify workflow patterns, trade-offs, and implications for generative video authoring tools.

4.2 Related Work

Our work is situated at the intersection of three research areas: video generation models, novel interfaces for video creation, and HCI research on dynamic documents.

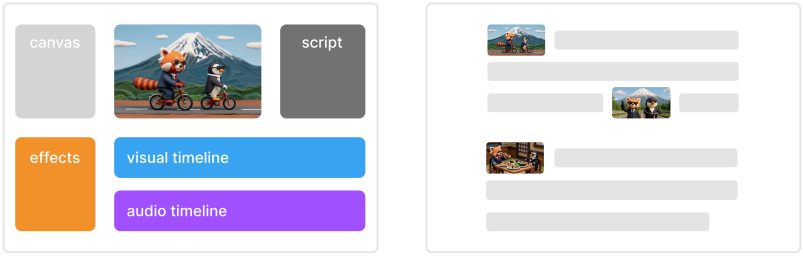
4.2.1 Generative Models as Individual Video Generators

Recent advances in generative AI have produced models capable of synthesizing high-fidelity video clips from text descriptions. Commercially available systems like Runway [113], Pika [103], Kling [70] and Google's Veo [50] demonstrate increasingly powerful capabilities. The primary interaction model for these tools is to write a text prompt to generate a single, short, fixed-length video. These prompts often require a structured format, specifying elements like the subject, action, context, and desired visual style [50]. More advanced features allow for greater control, such as using reference images to guide generation or maintain character consistency [113], and using start and end frames to define transitions [86, 50, 103].

However, the focus of these models remains on the generation of *individual, short clips*, typically lasting only a few seconds, *e.g.*8 seconds for Veo 3. There is limited to no support for structuring these shots into a longer, coherent narrative. Users must manually assemble the generated clips in a separate editing environment. Doki builds upon the capabilities

of such models. It treats them as the supporting engine while contributing a higher-level authoring interface that shifts the interaction from generating isolated clips to authoring a complete video story.

4.2.2 Novel Video Authoring Interfaces



(a) "Bento Box" like interfaces (b) *Doki*'s approach

Figure 4.1: A comparison of interface paradigms. (a) "Bento Box" style interfaces distribute authoring across multiple, separate representations. (b) Doki's approach uses a text-native canonical representation where the document serves as the primary interface.

A growing body of HCI research explores AI-powered interfaces for video authoring beyond traditional non-linear editors. Videorigami, for example, organizes authoring around multiple *compositional structures* (*e.g.*canvas, script, storyboard, timeline) and synchronizes them to support human–AI co-creation [23]. LAVE augments editing with AI agent and language annotations of footage [138]. ExpressEdit lets creators issue multimodal edit intents by combining natural language and on-frame sketching [131]. While these systems explore new interaction modalities, they typically distribute authoring across multiple panes – what we refer to as a "bento box" interface (Figure 4.1a) – which can incur split-attention costs when creators must reconcile multiple views and formats [26].

Doki embraces a contrasting design philosophy aimed at consolidating all authoring steps into a single representation.

Subtractive vs. additive workflows. Due to the nature of video capture prior to generative video models, most existing tools adopt *subtractive* workflows: creators begin with pre-recorded footage (or a library) and remove, trim, and re-order material. Examples include QuickCut [134]; transcript-aligned manipulation of talking-head footage [46]; and multimodal edit commands that operate on existing clips [131]. Some systems explore *additive* workflows where the tool synthesizes assets from text, such as Doc2Video's document-to-talking-head prototyping [29], but their system is limited to specific shot types rather than authoring complete, multi-shot stories. Doki takes an additive stance end-to-end: text is the substrate for *generating* shots and assets, not only selecting or trimming them.

Transcript-based editing. Transcript-centered tools let users edit a video by editing its text transcript, with changes reflected on the timeline [36, 111, 62, 29]. This interaction is especially effective for dialogue-driven media such as interviews and podcasts, where transcript and timeline are tightly aligned. Other systems use text to guide the assembly of existing footage: Write-A-Video [139] selects shots that semantically match an edited paragraph, while Crosscast and Crosspower augment audio with retrieved images [148, 146]. Doki complements these approaches by extending the role of text from mirroring transcript to *generating the full video* including audio tracks and video frames. Rather than treating text as just a proxy to transcript/speech tracks, Doki uses text to create and manage the visuals—and associated audio—end-to-end.

4.2.3 Rich-Text Editing and Dynamic Documents

Doki's interface design is inspired by research on dynamic documents [2, 64, 63, 78, 135]. Unlike applications that enforce predefined feature sets and rigid task boundaries, dynamic documents are designed as permissive mediums. Much like writing on paper, they do not impose a strict schema, allowing users to capture ideas in freeform ways. This flexibility supports a workflow of "gradual enrichment" [64], where users can begin with unstructured text

and progressively add structure and computational behavior as their needs evolve. Research prototypes such as Potluck [64] and Embark [63] demonstrate this principle by allowing notes to evolve into personal software or interactive trip plans.

Commercial tools have adopted related approaches. Notion [95] provides a block-based workspace where users can draft in natural text and later enrich content. OpenAI's Canvas mode [97] illustrates a similar trajectory. It extends chat-based interaction into an editable document environment where users can directly edit and request inline assistance.

Doki builds on these precedents but extends them into the video authoring domain. It treats video creation as the authoring of dynamic documents, that can be read as narratives and executed as a script, supporting a more flexible authoring process than existing video tools.

4.3 How People Create Generative Videos Today

We examined current end-to-end workflows for creating generative videos and describe motivating scenarios for Doki. Our analysis drew on public creators' generative AI video tutorials on YouTube and X [68, 1, 118, 125], and on observations of creators within our organization.

4.3.1 Motivating Scenarios

4.3.1.1 Creator A: Asset-First

Creator A anchored storytelling in asset construction. She typically spent 3-5 hours creating characters with text-to-image models and used the results as reference frames for consistency across a project. Yet, consistency remained elusive even with reference images. As she explained, *"I actually had two different octopus character reference frames that I used in GPT Image when I wanted more or less exhausted."* Because outputs often diverged, she repeatedly checked generation outputs and manually adjusted assets. After preparing characters and

environments, she created storyboards, generated short clips, and finally imported everything into a separate timeline editor to add voice and music. Her workflow required constant switching between asset creation tools, storyboard generators, video models, and editing software. Each step relied on a different view or interface, forcing repeated translation between formats.

4.3.1.2 Creator B: Script-First

Creator B started with a complete script – often drafted with a large language model – and decomposed it into scenes. Each scene was expanded into a detailed prompt specifying setting, character, tone, and cinematographic parameters. Because current video models lacked context awareness, every shot required re-describing all elements to preserve consistency. He explained, *"I would re-describe the setting, the character, and the tone every time."* This led to long and repetitive prompts, rendering small edits nearly impossible. Once clips were generated, they were reviewed and assembled in a video editor. This process prioritized prompt construction over narrative development.

4.3.1.3 Creator C: Iterative-Exploratory

Creator C worked in a highly exploratory style. Rather than preparing a script or asset library, he generated clips rapidly and used them to guide the next step. This improvisational process enabled serendipitous discoveries but became fragile during revision. For example, shifting the story setting from Tokyo to Barcelona required revisiting every prompt and regenerating all clips. As with others, he then manually reassembles on a timeline, further slowing iteration.

4.3.1.4 Recurring Challenges in Generative Video Authoring

Although creators approached generative video authoring very differently, many encountered the same set of challenges:

C1. Fragmented Tools and Formats: Even for a simple short video, many creators end up using three to five different tools — for example, a text editor for scripts, an image generator for visuals, a video model for clips, audio tools for narration or music, and a non-linear editor to pull it all together. The core elements of a project — scripts, prompts, images, audio, and edits — often live in separate formats across these tools.

This fragmentation can lead to constant context-switching, breaking creative flow and making it harder to maintain momentum. It also makes it difficult to get a clear overview of the project or understand how one change affects others. This raises the barrier to entry and slows iteration.

C2. Prompt Over Storytelling: Creators often spend disproportionate time crafting and tweaking prompts, which shifts focus away from narrative development and toward prompt engineering. For example, a 30-shot video might require 30 separate, verbose prompts – often managed manually in a list. Building up a complex prompt is itself difficult, and even once crafted, it is hard to control or manage across shots.

As a result, the workflow becomes less about shaping a story and more about wrestling with prompts, leaving creators with less bandwidth for the higher-level creative choices that should drive their work.

Polished rewrite (research-style, clear but precise)

C3. Consistency and Coherence: Maintaining visual and narrative consistency remains difficult. While recent models such as Nano Banana [45] and Flux Kontext [10] can generate consistent images from references, we argue that true coherence is as much an authoring and representation problem as it is a modeling problem. Conditioning on a few reference images does not provide the structural backbone needed to preserve assets, styles, and relationships

across an extended narrative. Without a representation that defines elements, propagates them through time, and manages context automatically, consistency will inevitably drift as projects grow.

4.3.2 Design Principles

Our analysis of existing workflows reveals a clear opportunity for a new authoring paradigm for generative videos. We propose four core design principles that guide the creation of Doki:

D1. Make text the central medium for authoring. C2 shows how prompt engineering can pull attention away from storytelling. Doki addresses this by making text the primary medium for authoring: prompts are embedded directly into the narrative, so the creative process feels more like writing a story than managing a command list.

Text has several advantages as a medium: it is familiar and accessible for humans, the native input modality for generative models, and inherently flexible and free-form. Because both humans and AI operate in the same medium, revisions and suggestions are immediately understandable: for example, AI-generated revisions are immediately understandable and can be refined in-place by the user, enabling fast and straightforward iteration.

D2. Consolidate video authoring in a single representation. In traditional workflows, creators switch between multiple views of the same project, which fragments the process and increases cognitive load (C1). Doki instead unifies scripts, prompts, visuals, audio, and timelines into a single canonical text representation. This representation functions both as the editable document that authors work with and as the underlying schema for the video itself. While not a full replacement for professional toolchains, this unified form enables many end-to-end creative tasks without leaving the document.

D3. Preserve consistency through parameterization. As projects scale, maintaining coherence across characters, styles, and scenes becomes increasingly difficult (C2, C3). Doki addresses this with a system of parameterized definitions that ensures elements remain

consistent while still allowing flexible variation. In addition, Doki automatically retrieves relevant references from earlier scenes so that continuity is preserved across the narrative. Together, these mechanisms make it easier to construct complex scenes, reuse components, and extend projects into longer, more cohesive stories.

D4. Keep interaction simple. Traditional non-linear editors are powerful but can be overwhelming (C1). Doki offers a simple editing environment with a lightweight slash menu for generating shots, defining assets, and inserting audio. Authors can write, generate, and edit within the paragraph, while using inline previews to review and trim.

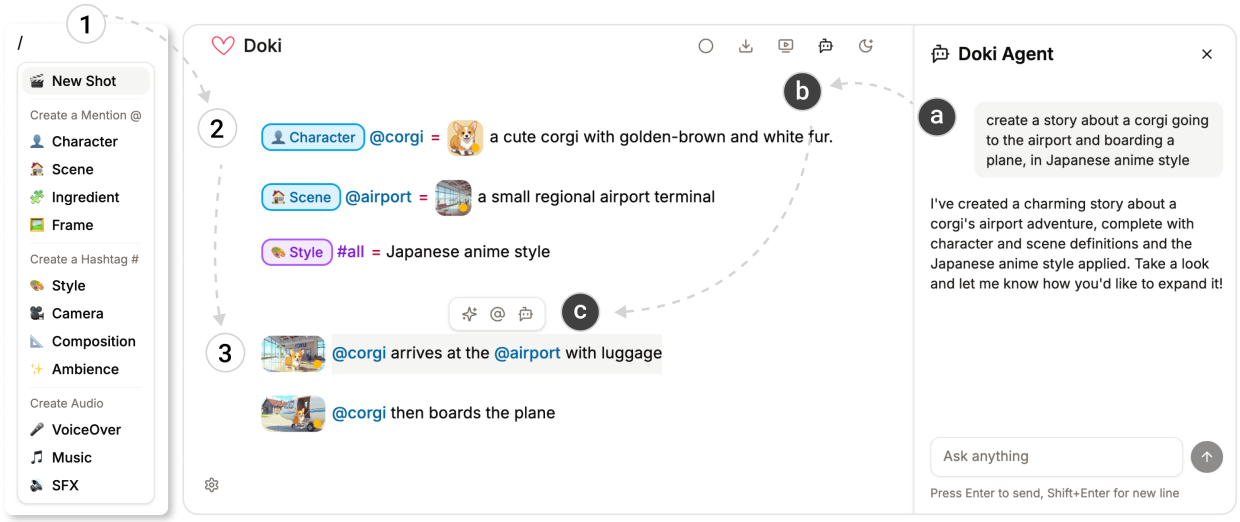


Figure 4.2: Two basic example workflows in Doki. Alice: (1) define assets and shots with slash commands → (2) write story and generate previews → (3) create video shots; Bob: (a) prompt the sidebar agent for a draft → (b) review the AI-generated draft → (c) refine with inline AI agent.

4.4 Doki

In this section, we first show example workflows of using Doki, then describe the structured text representation. Finally, we walk through each component of the system.

4.4.1 Example Workflows

Two creators produce the same short animation in different ways (Figure 4.2). The story follows a cute little corgi embarking on an adventure.

4.4.1.1 Alice: Writing a Video from Scratch

Alice prefers a hands-on, deliberate process. Starting with a blank text editor, she first defines the main assets for her story using slash commands. She types `/` and selects 👤 Character from the menu to create a definition block. Here, she names the character @corgi and provides a textual description: "a cute corgi with golden-brown and white fur". Similarly, she defines the setting with `/` → 🏠 Scene, naming it @airport and describing it as "a small regional airport terminal". To ensure a consistent visual aesthetic, she creates a global style definition using `/` → 🎨 Style, naming it #all and adding her description. For each definition, Alice generates a preview image. This allows her to check the visual interpretation and use the image as a reference to maintain consistency.

Next, Alice writes the storyline directly in the document. She structures it in paragraphs each corresponding to a video clip. For each clip, she creates a shot by selecting new shot from the slash menu (`/` → 🎬 New Shot). She writes the first shot: @corgi arrives at the @airport with luggage, and the second shot: @corgi then boards the plane. She then generates an image preview for each shot, following by the video. Once all shots are ready, she can export the video or view it in the built-in player.

4.4.1.2 Bob: Directing AI agents

In contrast, Bob prefers to leverage an AI agent workflow. Instead of starting from scratch, Bob uses the sidebar AI agent to generate the initial script. He provides a one-liner prompt: "a story about a corgi going to the airport and boarding a plane, in Japanese anime style". The agent generates a complete document structure, including definitions for the character, scene,

and style, along with two paragraphs of text for the initial shots.

To refine the agent's output, Bob can manually edit the document and the description of @corgi. Or, he can invoke the inline AI agent. For example, he selects the first paragraph and uses the following prompt: "add some background music here". The AI agent understands the context and appends this music description to the paragraph.

This approach also extends to revisions. Alice can manually change the character from a @corgi to a @cat, and Doki will regenerate the shots based on those specific changes. And Bob can use the sidebar agent for an agentic revision. He can prompt it to "adapt the story to happen in a New York City with a cat as the main character". The agent then revises the entire document, updating all definitions and shots to fit the new theme.

The examples above only demonstrate the basic uage of Doki. In real-world use, participants employ more intricate and blended workflows, which we examine in section 4.5.

4.4.2 A Structured Text Representation for Generative Videos

Document as Video, Paragraph as Sequence, Sentence as Shot At its core, Doki adopts a three-level hierarchical structure that draws inspiration from video production: the document as a whole maps to a video, paragraphs map to sequences, and sentences map to individual shots. This mirrors both the logic of filmmaking and the way humans naturally structure writings.

At the highest level, Doki conceptualizes the entire document as a video project. The arrangement of text determines the structure of the resulting video.

Within this framework, each paragraph corresponds to a sequence. Just as sequences in film bring together multiple shots to form a coherent scene, paragraphs gather sentences around a continuous narrative. A new paragraph often marks a shift in focus, such as transitioning from one topic to another, which parallels the cinematic logic of moving from one scene to the next.

At the most granular level, sentences maps to shots. In filmmaking, a shot is a single, uninterrupted take. In Doki, authors insert a shot preview element into the document to begin a new shot. All the sentences that follow this marker, up to the next shot or paragraph break, are treated as the textual description of that shot.

This three-level hierarchy resonates with how writers already think and compose: when beginning a new topic, they start a new paragraph; when articulating a small unit of idea, they write a sentence. Doki aligns these practices with the structure of video production.

Parametrization and Context Beyond the basic document structure, Doki enriches this text representation through parametrization and context handling. We introduce a definition system that allows authors to define and reuse key elements consistently across the document. It also supports automatic reference resolution, ensuring continuity across sequences and shots. Implementation details of these mechanisms are elaborated in the following sections.

4.4.3 Doki System

The Doki interface is intentionally minimal. Upon opening, users see a blank document with only a few controls to learn. In the top right corner, the interface provides a status indicator, download, video player, AI agent, and a dark mode toggle. In the bottom left corner, a settings button is available. The central interaction space is the document itself.

4.4.3.1 Slash Menu

When users type a slash `/` (Figure 4.2), a menu appears that provides quick access to key creation tools. From this menu, they can start a new video shot, create reusable definitions like characters and scenes, or add audio to their project.

4.4.3.2 Shots

In Doki, *shot* is the fundamental unit of a visual story. Each shot is embedded directly in the text as an inline element, tightly woven into the flow of the narrative. This design allows writing and visuals to coexist in the same space.

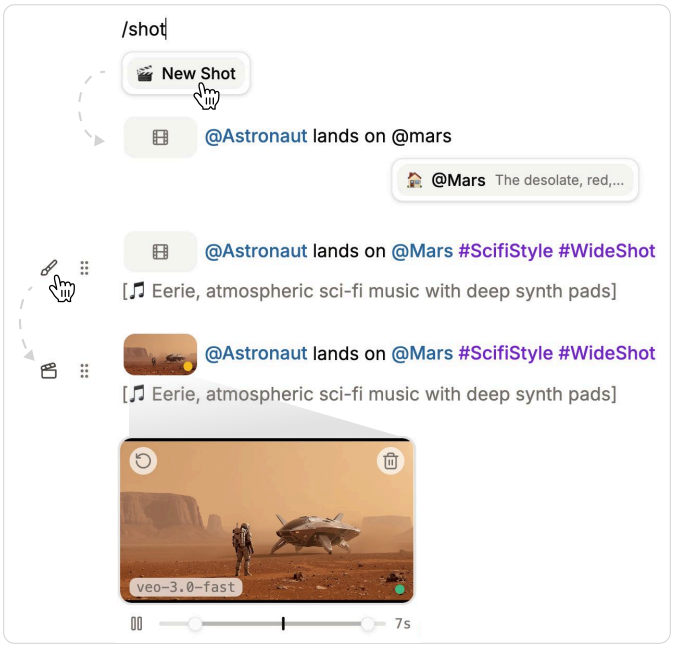


Figure 4.3: Creating a shot in Doki. A new shot is inserted inline with a slash command, and the description that follows serves as its prompt. The system first generates a preview image, which can then be turned into a video clip. Users can click to expand them for playback and additional controls.

Creating a Shot. To create a shot (Figure 4.3), a user types `/` → 🎬 New Shot, which inserts a shot preview node 🎞. The description that follows the node serves as the raw prompt.

Shot Sequences. As we described in subsection 4.4.2, paragraphs in Doki function as analogues to sequences. Users may insert multiple shots within a single paragraph (Figure 4.4).

In this case, later shots use the image generated by the preceding shot as contextual input for the model. This supports continuity across a sequence. For example, writing:

🎞 @Pandi riding a bike. 🎞 #CloseUp on @Pandi

produces two connected shots. The second builds on the first by applying a close-up shot on the character in the same scene.

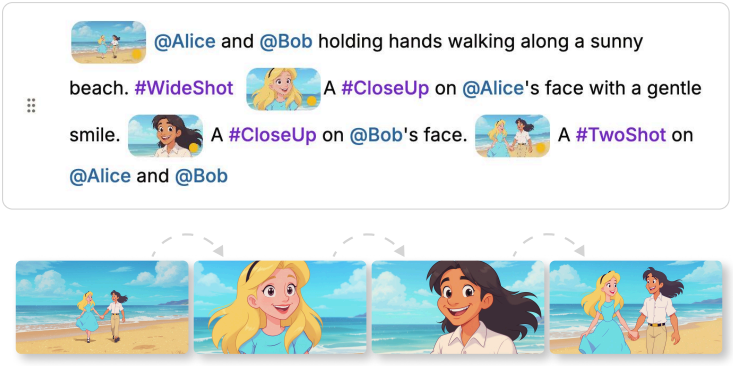


Figure 4.4: Writing consecutive shots within a single paragraph. Later shots inherit context from earlier ones, enabling continuity across a sequence. We achieve strong consistency between shots without repetitive context description in the Doki document.

Shot Variants. Users can generate multiple variants (Figure 4.6) of a shot by right-clicking and selecting Add Variants. For example:

🎞 🎞 🎞 @Pandi riding a bike.

This action creates three alternative previews using the same prompt. The first is selected by default, while the others appear semi-transparent as backups. Users can drag any variant to the front, establishing it as the active version.

Generating Image and Video. Each shot follows a staged generation pipeline: text → image → video. The system first uses the shot's description to generate a preview image,

which then will be used as the first frame for video generation. The inline shot preview node will be updated accordingly as this progress proceeds. This staged pipeline offers authors control and aligns with existing creative workflows. Our formative studies revealed that creators typically prefer generating an image before producing video, as it provides an opportunity to preview visual direction before committing to a full video. Cost differences reinforce this preference: generating video costs $3.20 per clip (Veo 3), compared to $0.04 per image generation (Imagen 4). For shots in the same paragraph, Doki processes each shot in order, using prior images as contextual references for subsequent ones (details in subsubsection 4.4.3.5).

Shot Status. Each shot displays its current status through a small circular indicator in the bottom-right corner. The indicator communicates whether a shot is generating an image (breathing yellow), image ready (static yellow), generating a video (breathing green), video ready (static green), or outdated (red). A shot becomes outdated when its prompt or any referenced definition changes.

Expanded Shot. While shots are compact inline elements by default, they can be expanded with a click. This expanded view (Figure 4.3) provides additional controls, allowing users to trim videos, delete, or regenerate. Model information is also displayed.

4.4.3.3 Parametrized Definitions

Definitions are reusable parametrized building blocks in Doki. They allow creators to specify characters, scenes, styles, and other elements once and reference them throughout the document.

Types of Definitions. Doki supports two categories of definitions: @Mentions and #Hashtags. Mentions define elements of the story, while hashtags specify qualities that shape how

the story appears. Mentions act like nouns, while hashtags act like adjectives.

@Mentions include:

- 👤 Character — a character in the story (*e.g.*@Pandi, a little red panda wearing a suit and sunglasses).
- 🏠 Scene — a location or setting (*e.g.*@subway, a crowded subway cabin in Tokyo).
- 🧩 Ingredient— an object or prop (*e.g.*@sushiBox, a wooden bento box filled with sushi).
- 🖼️ Frame — a reference frame (*e.g.*@PandiSubway, a frame of @Pandi in the @subway).

#Hashtags include:

- 🎨 Style — the overall visual aesthetic (*e.g.*#filmnoir, #claymation, #anime).
- 🎥 Camera — cinematic movements (*e.g.*#closeUp on @Pandi, #Pan across @TokyoTower).
- 📐 Composition — spatial arrangement or framing (*e.g.*#twoShot, #wideShot).
- ✨ Ambience — mood, color, or lighting (*e.g.*#sunsetGlow, #neonLights, #misty).

Doki also provides a built-in cinematography library with professional terms such as #CloseUp, #Pan, and #Dolly, making it easy to apply established film techniques.

Creating Definitions. To create a new definition, users type the slash `/` to reveal the menu, then select the definition type. This inserts a tag indicating the definition type, followed by the appropriate prefix symbol (either @ for mentions or # for hashtags). Immediately after the prefix, users first provide a name for the definition, then enter an equals sign = and a natural language description (Figure 4.5). For example:

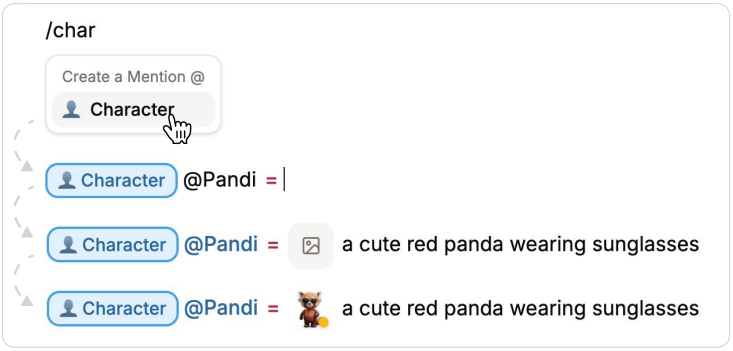


Figure 4.5: Creating definitions in Doki. Users open the command menu with a `/`, select a type, and provide a name and description. Optionally they can add a visual definition for even better consistency.

Character @Pandi = a little red panda

Definitions can be combined and nested. For instance, one character definition may include references to other characters, props, or styles. This allows users to incrementally build more complex concepts from simpler ones, while maintaining reusability across the document.

Visual Definitions. In addition to textual descriptions, users may attach a visual preview to any definition. To do so, insert a shot using the slash command `/` → 🎬 New Shot immediately after the equals sign (Figure 4.5):

Character @Pandi = a little red panda

If a visual is added, Doki uses it as the primary reference for all future generations involving that asset. If no visual is present, the system falls back to the text description. Similar to shots, visual definitions can have multiple variants (Figure 4.6). In addition, Doki also supports user-uploaded images for definitions. A user can upload a custom image by selecting Upload Image in the right-click context menu. The system automatically analyzes the image and generates a corresponding textual description using Gemini 2.5 Flash.

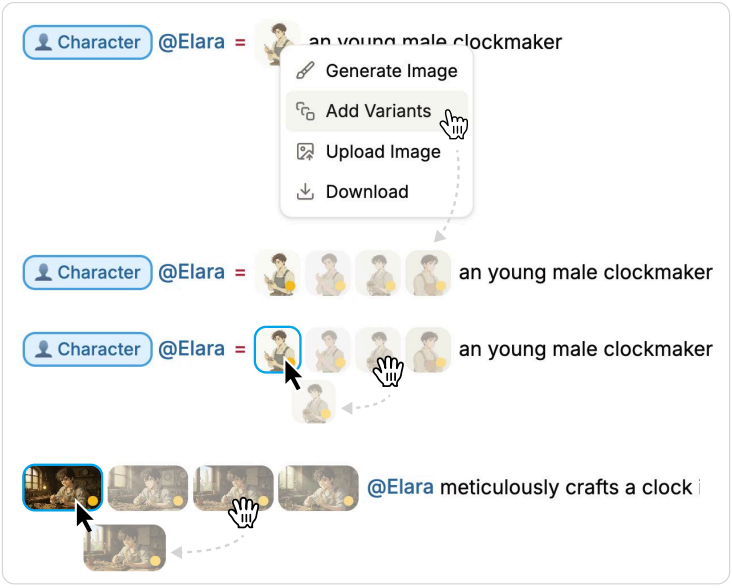


Figure 4.6: Adding variants to a definition and shot. The user selects Add Variants from the context menu by right-clicking on the image. Three variations will be added inline. The user can then select one variant by moving it to the first position.

Referencing Definitions Once a definition is created, it can be reused throughout the document. To reference a definition, the author types either @ or # followed by the name. Doki provides auto-suggestions as the user types, allowing quick selection from a list of existing definitions (Figure 4.3).

References are linked to their source definitions. Any change to a definition automatically propagates through the document. For example, if @Pandi is renamed to @Panda, every occurrence of that character updates instantly. Similarly, if the description of a definition is revised or a new visual is attached, all references will reflect the updated content.

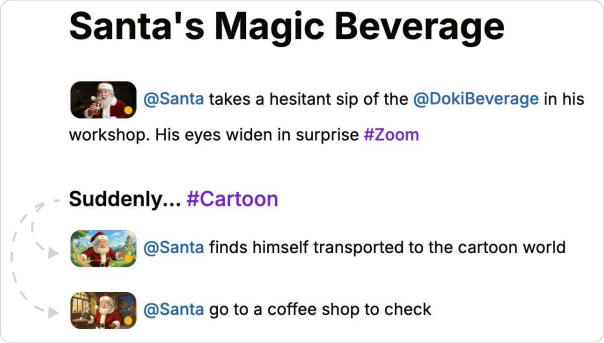


Figure 4.7: Authors can apply definitions globally (#all =) or scope them to specific sections via headings. In the example, a scoped definition (#Cartoon) transforms subsequent shots under the heading level into a consistent cartoon style.

4.4.3.4 Global and Scoped Definitions

As documents grow in length in Doki, authors often need to apply the same definition, such as a #Style or @Scene, across multiple paragraphs. Repeating these references manually can be tedious. To address this, Doki supports both *global* and *scoped* definitions that automatically propagate without requiring explicit referencing.

To define a global attribute, users assign the name "all" to any definition. For example, writing:

🎨 Style #all = Photorealistic

automatically applies the photorealistic style to the entire document. This syntax works for all types of definitions, including mentions and hashtags.

In addition to global definitions, Doki supports heading-level scoping. Documents often contain a structured hierarchy of headings, and Doki leverages this structure to control the scope of definitions. Placing a definition on a heading applies that definition to all content within the corresponding heading level (Figure 4.7). This design gives authors predictable control: for example, overall stylistic rules can be set once, and specific deviations can be introduced at finer levels without redundancy.

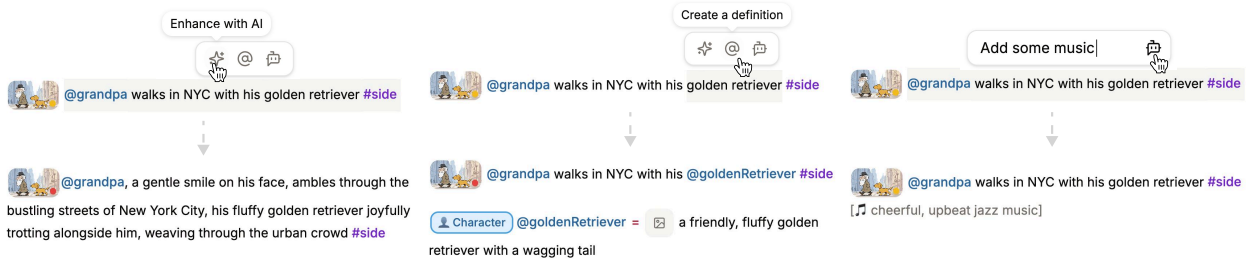
4.4.3.5 Paragraph Handles

When hover over a paragraph, up to two buttons may appear on the left. On the right, the handle ⠿ allows users to reorder the paragraph or its associated shot sequence. The left-side handles communicate available generation options. If no shot is eligible for generation, the handle remains hidden. When the system detects that preview images can be produced, it displays a paintbrush icon 🖌. If the paragraph is ready for video generation, the system instead shows a clapperboard icon 🎬. Clicking the paintbrush triggers image generation for the paragraph. Clicking the clapperboard initiates video generation. The system does not regenerate all assets indiscriminately. Instead, it uses an algorithm to determine what needs to be generated and in what order.

4.4.3.6 Audio

Users can add audio to their shots through the slash command /. From this menu, they may insert 🎤 Speech, 🎵 Music, or 🎧 SFX. Authoring audio is via text in square brackets. The system interprets any text enclosed in brackets as audio. For clarity, the notation appears in gray and can be mixed with normal text.

When audio notations are included, they are passed to video generation models that support synchronized audio, such as Veo 3 and Veo 3 fast. If no audio is specified, the system produces no/minimal sound effects.



(a) Enhance (b) Create Definition (c) Custom Request

Figure 4.8: Three types of inline agent actions. (a) "Enhance" increases the descriptiveness of the selected text. (b) "Create Definition" converts the selection into a reusable definition. (c) "Custom Request" applies an in-context edit based on user instruction.

4.4.3.7 AI Agents

Doki integrates two AI agents that provide complementary forms of assistance: the Sidebar Agent, which operates as a turn-based conversational assistant, and the Inline Agent, which supports direct, text-level edits within the editor. Both agents can perform small adjustments as well as large, multi-step edits. All edits are applied directly to the document and highlighted visually.

Sidebar Agent The Sidebar Agent can be opened from the main header and interacts with users through a conversational interface. Because the agent has full context of the document, it can carry out edits that span text, definitions, and generated media.

For example, it can support requests like:

- Expanding concepts. *e.g.*"Make the #goldenHour definition more detailed."

- Inserting visual references. *e.g.*"Add a visual reference to the @castle definition."
- Refining pacing and tone. *e.g.*"Fix the pacing of the video. Make the first two shots more dramatic and add a climactic final shot."

Inline Agent The Inline Agent supports immediate, in-context editing within the document. It appears only when text is selected, showing a bubble menu with three options: Enhance, Create Definition, and Custom Request (Figure 4.8): *Enhance* increases the descriptiveness of the selected text. *Create Definition* converts the selection into a reusable definition and replaces the text with that reference. *Custom Request* opens a small input field for user instructions.

With the inline agent, authors can begin with rough drafts or loosely organized ideas and progressively refine the drafts into structured narratives. For example, a writer might draft the sentence "A young boy runs across the courtyard." At first, the boy is an incidental figure, but the author later decides to establish him as a recurring character. By selecting "A young boy" and clicking @ to turn it into a persistent @Character definition. Then, the author can issue a custom request, such as "make him wear a hat", to expand the reference to add more details.

4.4.3.8 Other Notable Features

Doki also provides a set of other features to support the authoring process in the header and the settings panel.

• *Status Indicator*: A small circular icon communicates the current state of the system.

• *Export*: Users can export their work at any time. The download menu supports three formats: the full rendered video, a zip file of all individual shots, and a JSON file containing the project's structured data. This enables seamless transfer across devices and external editing tools.

• *Video Player*: A built-in video player allows users to preview their generated content without leaving the application. The player can be toggled open or closed and provides basic playback controls for reviewing the full sequence. It also supports shot-level navigation. As playback advances from one shot to the next, the system automatically scrolls the document to the corresponding section.

• *Settings Panel*: The lower-left corner includes a settings menu for project-level actions. "Load" allows users to open a saved project from a JSON file. "New" starts a fresh project. Users may select among supported AI models for text, image, and video generation.

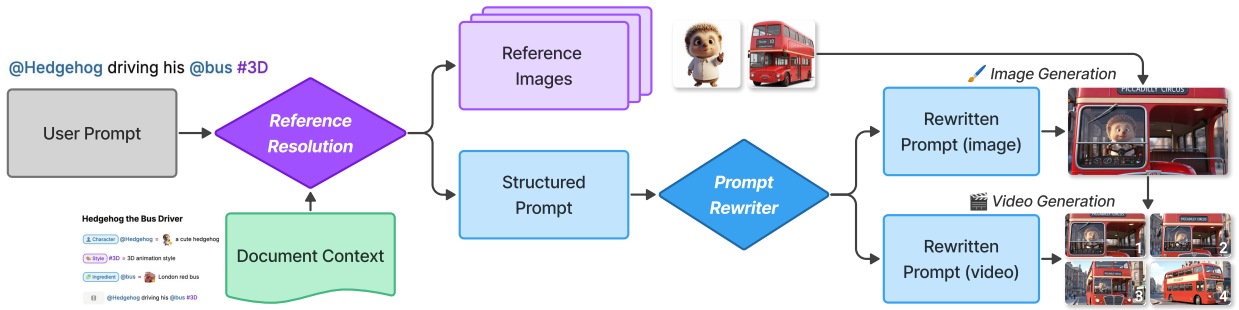


Figure 4.9: Doki's shot generation pipeline. First, the user's raw prompt and the document context is passed to reference resolution module to create a structured prompt and gather relevant visual reference images from the definitions. Then, the prompt rewriter rewrites and polishes this structured prompt for image and video generation. Doki first produces a static preview image and then uses that image as a the first frame to generate the final video clip.

4.4.4 Doki's Shot Generation Pipeline

To transform a user's text input into a video shot, Doki maintains three distinct types of prompts throughout this process (Figure 4.9). The first is the *user prompt*, which corresponds to the text entered directly into the editor. The second is the *structured prompt*, which augments the user input with relevant references to definitions. The third is the

rewritten prompt, which refines and polishes the structured prompt to optimize performance for downstream image and video generation models.

4.4.4.1 Structured Prompt

Once the system receives the user prompt, it parses the text and resolves references to produce a structured prompt through a tree structure. This step grounds each reference in the prompt with its corresponding definition. For example, for the user prompt: @Hedgehog driving his @bus #3D, the system transforms it into:

```
DESCRIPTION: @Hedgehog driving his @bus #3D
DEFINITIONS:
Character @Hedgehog = a cute hedgehog in a white shirt
Ingredient @bus = London red bus
Style #3D = 3D animation style
```

Instead of relying on the user prompt, Doki uses structured prompts to maintain consistency across revisions. For instance, updating the description of @Hedgehog will automatically mark all shots involving that character as "outdated" (little red dot on bottom right), even if the user did not modify the shot text directly.

4.4.4.2 Rewritten Prompt

Doki then refines the structured prompt through a prompt rewriter. The structured prompt is accurate but often rigid and unnatural. The rewriter addresses this by producing a fluent natural language description optimized for generative models. In practice, the rewriter is implemented as a large language model prompted with carefully designed instructions. Prompts used for image and video rewriters are available in the supplementary materials.

4.4.4.3 Selecting Image References

Beyond resolving textual references and refining prompts, the system also retrieves relevant visual references to guide shot generation. When a definition includes an associated preview image, the system incorporates that image as a reference. In addition, when a paragraph contains multiple shots, the model retrieves the image from preceding shot in the same paragraph. This mechanism preserves visual continuity across a sequence, as illustrated in Figure 4.4. In addition to rule-based strategies, the system uses a prompted LLM (full prompt in supplementary materials) to identify image references that are contextually relevant but not explicitly referred to. For instance, if a story introduces a location, shifts to another scene, and later returns to the original setting, Doki can infer that earlier images of the location should be referenced to maintain consistency.

4.4.4.4 Image and Video Generation

Once the system produces a rewritten prompt and retrieves reference images, it can now generate images and video for a shot in Doki. As described earlier, generation always begins with a static image preview that serves as a first frame, and then from there generate the video.

For image frame generation, the model combines the rewritten prompt with the retrieved reference images as context. By default, Doki uses the `Flux Kontext Pro` model for image generation, which takes both text and images (up to 4) as input. Video generation defaults to `Veo 3 Fast`.

4.4.5 Implementation

Doki is built with TypeScript and React for the frontend, TipTap for rich text editing, Zustand for state management, Node.js for backend, and FFmpeg for video processing.

4.5 Diary Study

We conducted a week-long mixed-methods diary study of Doki. We designed our study around three research questions:

RQ1. How do users perceive the benefits and limitations of Doki's text-native interface?

RQ2. What workflows do people employ when using Doki?

RQ3. How does prior experience with video editing and generative AI influence the ways they perceive and use Doki?

We chose a diary study over a lab study for three reasons. First, there is no established point of comparison for generative video authoring. Existing tools are designed for captured content and follow very different paradigms. Second, because generative video is new, most people have little prior experience, and a short session can only reveal initial impressions rather than deep insights. Third, Doki is a freeform system – we expected participants to adapt the system and develop their own workflows, which can only emerge with extended use. A diary study offered the best conditions to help us learn these evolving practices.

4.5.1 Participants

We recruited ten participants (6 female, 4 male; Avg. age=32.4, SD=9.9, range=[24, 57]) for a week-long diary study of Doki. Six were recruited from outside of our organization and four from within. Roles included product designers (2), animators (2), a filmmaker, a graphic designer, a software engineer, a content creator, a UX designer, and a program manager.

Participants reported different levels of experience and frequencies of video creation: 4 created videos daily, 3 weekly, and 3 a few times per year. Their videos spanned diverse contexts, such as professional films, social media content (*e.g.*TikTok, Reels, Shorts), academic/-

work projects, and personal vlogs. Participants used multiple video editing tools. Adobe Premiere Pro was most common (8/10), followed by iMovie (5) and CapCut (5). Three reported using specialized generative video platforms (3/10, e.g., Runway, Pika, Google Flow).

Participants' experience with generative AI tools also varied. For large language models, 5 reported daily use while 3 shared rarely using them. Image and video generation tools were less commonly used: 4 participants reported using image generation models at least once per week, and 2 reported using video generation models at least once per week. Notably, the participants who used video generation models regularly had professional roles closely tied to generative video creation: one is a filmmaker producing AI-driven video content; the other a product designer responsible for producing animation videos for their brand. Overall, seven participants were already using generative AI in their existing video creation workflows. For example, scriptwriting with ChatGPT, storyboard ideation with Midjourney, refining edits with Photoshop's AI features, etc. 3 participants reported minimal or no prior experience with AI tools – providing contrast in perspectives and practices.

4.5.2 Procedure

The study followed a three-phase structure:

Phase 1: Onboarding (60 minutes) Participants completed informed consent and a background questionnaire, received a guided tutorial to Doki's core concepts (e.g., shots, definitions, slash menu, agents) and completed a short open-ended creative task ("*Create a 30-second video about anything*") with support from one of the authors.

Phase 2: In-the-Wild Diary Study (5 days) Participants used Doki independently for five days. Each day, they were instructed to engage with the tool for at least 50 minutes, with the goal of producing 2-3 complete videos at the end of 5 days. Participants were encouraged to follow their natural workflow, allowing ideation, drafting, generation, and

iteration to unfold across days.

After each day, participants completed a brief survey. To capture an overall measure of participant satisfaction, we included a single-item rating question (1-10): "*Please rate your overall satisfaction with your Doki experience today*". Participants also reported their workflow strategies, identified moments of ease and difficulty, and listed features used (*e.g.*shot generation, definitions, AI agents, etc.). Each survey concluded with an optional submission of the day's project file and exported video.

In addition, we collected interaction logs spanning session duration; number of shots created or deleted; editing operations; words typed; use of mentions and hashtags, previews, and menus; generative activity (*e.g.*attempted/successful image or video generations, regeneration counts, total generation time); export behavior; AI assistant interactions; and cost metrics based on image and video generations.

Phase 3: Exit Interview (60 minutes) Participants completed a semi-structured remote interview. The first portion consisted of a retrospective think-aloud session [52], in which participants reviewed one of their projects, explaining creative decisions, feature use, and how the final video evolved. We also asked about overall experience, workflows, ownership, and learning. Participants also completed the System Usability Scale (SUS) [19] to provide standardized quantitative feedback.

Apparatus Participants accessed Doki remotely as a web app during the week-long study period. External participants were compensated $30/hour. Collected data included submitted videos, Doki documents, surveys, interaction logs, and interview recordings/transcripts.

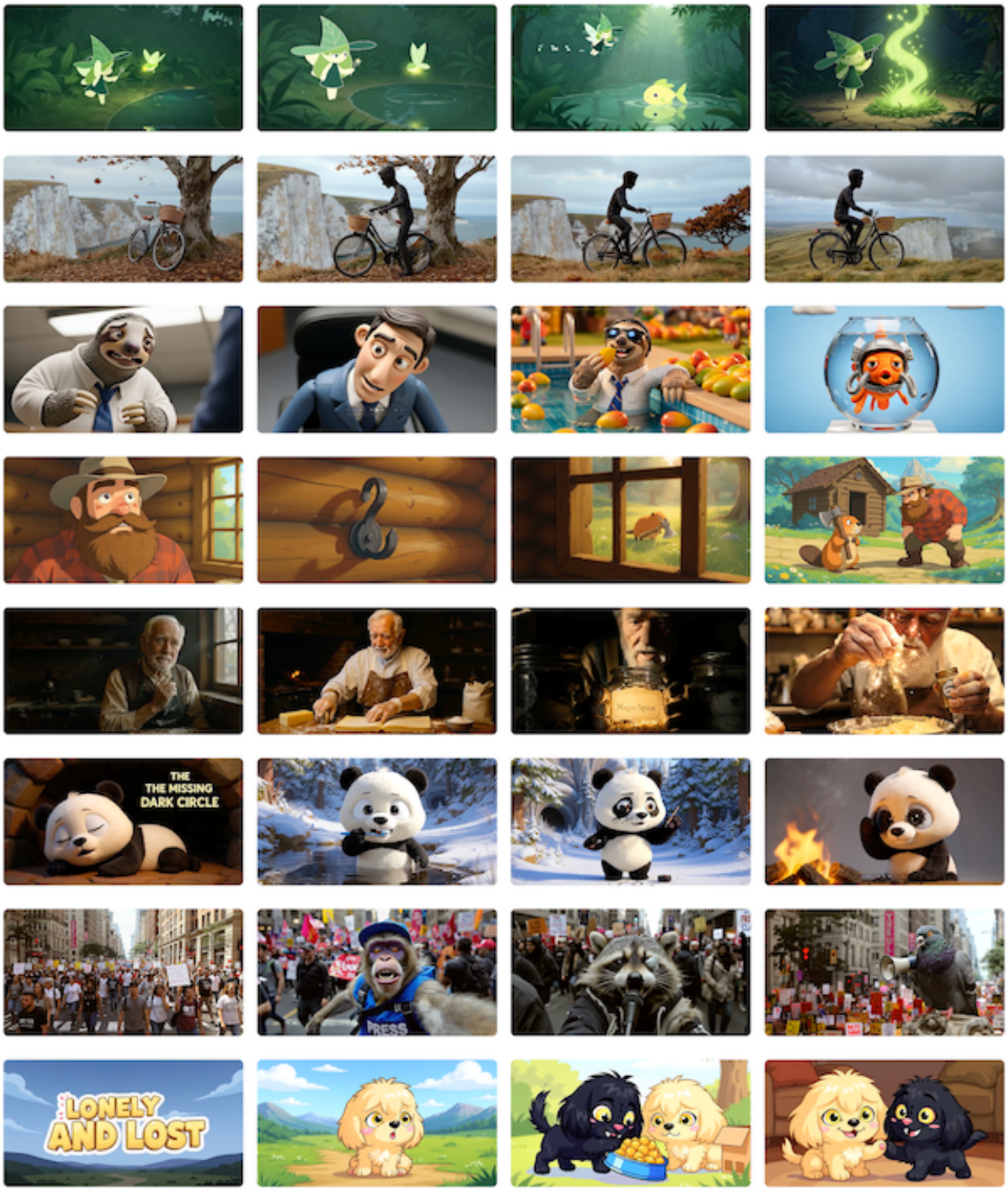


Figure 4.10: Example frames from videos created by participants in our diary study. The videos demonstrate the breadth of video types Doki supports. Each row shows a video through four representative keyframes.

4.6 Results

4.6.1 Doki Usage

4.6.1.1 Interface Usage

Excluding the onboarding and exit interviews, participants spent an average of 91.7 minutes per session (SD = 45.5; range: 40–217) editing with Doki, indicating engagement beyond the required 1 hour per day. On average, they generated 45.5 images (SD = 33.2) and 20.3 videos per session (SD = 15.8). Normalized by time, participants produced approximately 0.6 images and 0.3 videos per minute.

The most frequently used features were export (96% of sessions), shot generation (90%), and the video player (90%). Participants also made extensive use of defining hashtags (82%) and mentions (78%). Shot variants were generated in 76% of sessions, and audio was added in 68%. The chat sidebar AI agent was used in 74% of the sessions while the inline agent was used 50%. Heading-based styles were the least used feature, appearing in only 24% of sessions.

4.6.1.2 What Videos Did Participants Create with Doki?

Participants submitted 46 video files in total. The average duration was 67.14 seconds (SD = 35.97; range: 15.04–184.04). Most videos (31) fell within 30-90 seconds; Four were shorter than 30 seconds, and eleven exceeded 90 seconds.

Participants produced a variety of videos (Figure 4.10) that fell into five main categories: storytelling, instructional video, advertising, music, and experimental. Storytelling was the most common, with eight participants (P2, P3, P4, P5, P7, P8, P9, P10) creating original narratives such as fairy tales, pet stories, or character-driven animations. One participant (P1) focused on instructional formats, producing cooking videos and academic explainers. Advertising and promotional content appeared in three cases: P2 created two

claymation-style advertisement for Doki, P6 worked on a summer camp promotion video, and P8 produced a series of animal welfare videos tied to World Dog Day. Some participants explored less conventional directions. P8 experimented with music videos. Two participants also created experimental projects like a surreal animal protest (P2) and ASMR (P7).

While many participants used Doki for narrative storytelling, these projects demonstrate the breadth of video types Doki can support. We include selected videos authored by participants in the supplementary material.

Id	Role	Video Creation Freq.	GenAI Use (Text)	GenAI Use (Image)	GenAI Use (Video)	SUS
1	UX Designer	A few times a year	A few times a year	Monthly	A few times a year	87.5
2	Filmmaker	Daily	Daily	Daily	Daily	77.5
3	Product Designer	Weekly	Rarely or never	Rarely or never	Rarely or never	75.0
4	Software Engineer	A few times a year	Daily	Weekly	Monthly	90.0
5	Animator	Daily	Rarely or never	Rarely or never	Rarely or never	80.0
6	Program Manager	A few times a year	Daily	Monthly	A few times a year	72.5
7	Animator	Daily	Weekly	A few times a year	Rarely or never	62.5
8	Product Designer	Daily	Daily	Weekly	Weekly	87.5
9	Graphic Designer	Weekly	Monthly	Weekly	A few times a year	90.0
10	Content Creator	Weekly	Daily	A few times a year	Rarely or never	87.5

Table 4.1: System Usability Scale (SUS) scores and subscales (Usability and Learnability) across participant roles, video creation frequency, and generative AI tool use.

4.6.2 RQ1. How do users perceive the benefits and limitations of Doki's text-native interface?

4.6.2.1 Overall impression of Doki

Overall, participants reported a positive experience with Doki. The System Usability Scale (SUS) ratings (Table 4.1) collected during the exit interviews averaged 81.2 (SD = 8.9, median = 83.8). This score corresponds to the *Excellent* category defined by Bangor et

al. [8], placing Doki within the 90–95th percentile (benchmark average = 68). Participant ratings ranged from a 62.5 to 90.0, with all but one participant rating Doki at or above 72.5. These findings are consistent with daily survey measures of overall satisfaction, which had a mean of 7.48/10 (SD = 1.68, median = 8.0).

Beyond the quantitative measures, many participants described Doki as a robust system that generates high-quality visuals and is delightful to use. For example, P8 shared that "*This whole process makes me feel relaxed and gives me a great sense of accomplishment.*" Similarly, P1 remarked, "*Doki is a really great generative AI video tool, probably one of the best I've used so far.*"

4.6.2.2 Ease of use and learn.

Nine out of ten participants described Doki as very easy to use, highlighting its simple interface. They noted that the system required little effort to learn and that they could quickly grasp its core functionality. For example, P3 and P5 reported that common features such as the slash command, mention system, and hashtags mirrored those in platforms they already used daily. This familiarity helped them learn quickly and made the features easy to remember. P8 found Doki much simpler than her usual video creation workflow because it minimized context switches:

> "*Doki combines them together so I don't need to shift between different windows...it really saves me the window shifting time.*"

These qualitative findings align with the survey results. Quantitative analysis showed no significant change in participants' satisfaction scores across the study period (Repeated Measures ANOVA, p = 0.42). Participants consistently rated their experience positively from the start. As P4 noted, "*Almost immediately… I felt like I could use it right away.*" In the exit interviews, the System Usability Scale (SUS) subscale for learnability [77] yielded a median score of 81.2, which falls within the excellent range.

Because of Doki's simple interface, three participants mentioned that Doki made them *focus on storytelling*. Compared to tools like Flash or After Effects, P9 described Doki as "*10 times easier*," and he could focus on the idea and content rather than symbol animation or movement. As P5 put it, "*I think the interface is perfect...the whole interface is very simple and clean. So I can focused on creating my story.*"

4.6.2.3 Fast to go from idea to content.

Eight participants emphasized how Doki accelerated the process of moving from a rough idea to deliverable content. It enables participants to bypass many of the slow, manual steps in traditional workflows. P1, a former professional documentary filmmaker, contrasted Doki with traditional production workflows:

> "*For like a quick project, I think with my previous workflow that's just not possible because I have to go out to shoot, I have to spend a lot of time on editing. But with Doki I can just easily prompt the system to generate a very great first draft.*"

Several participants mentioned that this efficiency was especially valuable for fast, lightweight projects where perfect visual fidelity was not the priority. P5 reported: "*So on average it will be 15 minutes to create about one minute or longer video.*" Similarly, during the diary study, P8 produced a series of five 30-second videos for World Dog Day, in roughly ten minutes each – an output that would have been unthinkable in their prior workflow. Participants consistently noted that integrating an AI agent within the document interface was essential to this speed. As P2 described, "*It can create short level breakdowns. It can auto format the whole scenario. It can do anything for you.*"

4.6.2.4 Document representation fosters comprehension

Four participants explained how Doki's document-based representation not only accelerated content creation but also enhanced their comprehension of narrative flow and structure.

Unlike traditional non-linear editors (NLEs), which primarily rely on timelines, Doki offered a holistic, text-based overview of their projects. This shift was seen as particularly valuable in helping participants visualize their films in a semantic way. P2, a professional filmmaker, highlighted how this representation provided an immediate sense of coherence across the film:

> "*Doki sets itself apart in how it gives you a sense of how your film will look. You can see the flow of your scenes, how your film is looking scene by scene and shot by shot.*"

By contrast, P3 noted that conventional NLEs often lack such an outline, making it harder to conceptualize story development beyond technical editing: "*It (NLEs) doesn't have a nice visual outline for writing or editing your videos or cropping.*" Similarly, P6 explained that the document structure enhanced her understanding of video composition: "*I feel like it was helpful for me to understand the structure of [the] doc. It's pretty intuitive.*"

In addition, participants mentioned that the text representation also made it very easy for them to view and understand AI's changes, and even for them to learn how to better use Doki (P4):

> "*The notebook interface with text as the driving element is clever. You can learn from how it generates music or voice pieces, even if you wouldn't use it directly.*"

4.6.2.5 Parametrization makes story building easier

Six participants emphasized how Doki's parametrization design helped them build their stories. For example, by defining reusable elements such as characters, scenes, and styles, they could begin projects with a strong foundation even when the narrative was not yet fully formed. As P6 explained, "*Building reusable characters and ingredients really made sense as a lot of times I have some clarity into what/who I want to use but not the story itself.*"

Participants also described how these definitions supported coherence across their work. P10 highlighted that parametrization reduced randomness in generative outputs. "*I think this tool is really cool because you can define a character and scene so it would not be like very randomly generating the video.*" Similarly, P1 noted that Doki itself is not just a video tool but also a really good tool for prompt engineering and crafting the visuals, even just a single shot. He described that the definition system make it easier for him to build up complex scenes from smaller components:

> "*I don't have to write repeated keywords just to keep this consistent. I can just use tags and add each character so and they will stay.*"

He further envisioned that in the future, professional creators could design and share parametrized styles or camera movements as creative assets, and make them available through simple tags for others to use.

Finally, participants appreciated how parametrization integrated naturally into Doki's text-native representation. P5 noted how features like mention, hashtag and audio functioned seamlessly within the document: "*...mention and hashtag and create audio. It's all in the same logic thing. It's very easy to remember and learn.*"

4.6.2.6 Model control is limited.

Nine participants noted that Doki's outputs often failed to match their prompts, requiring repeated regeneration to achieve acceptable results. P1 described persistent artifacts: "*A lot of videos include nonsensical text that I couldn't remove, even when I explicitly prompted it.*" Such errors forced participants into trial-and-error workflows. As P10 explained, "*I need to regenerate image again and again...* " P8 also noted visual glitches such as "*dogs disappearing mid-run and people brushing the air instead of the animals.*" Several expressed frustration that the text-only interface constrained their ability to realize specific visual goals. P9 noted, "*Doki is not good at when you want specific frame compositions.*" This challenge

was especially pronounced for users with *strong visual references*. During the study, P10 attempted to create a video from her storybook. Because she already had a clear mental image of the visuals and compositions, she found it difficult to reproduce them through text prompts. She explained, "*If you want to create your personal story, sometimes the video could not reflect everything that you want. Maybe just achieve 80% of your imagination.*"

4.6.2.7 Hard to work with music.

Participants noted constraints in how Doki handles time-based media. Because the system is organized around a paragraph and document structure, it is difficult to add audio that spans multiple paragraphs or begins asynchronously. P8, who attempted a music video, found it hard to align visuals with audio.

4.6.3 RQ2. What workflows do people employ when using Doki?

4.6.3.1 A lot of variety in how people start.

Participants entered Doki from different starting points, reflecting varied creative practices. Some began with a complete script that they had written beforehand. For example, P1 noted: "*I already have a text version of that recipe, so I just use the chatbot, paste the whole recipe and ask [it] to generate all the ingredients I may need.*"

Others would first defining characters, scenes, or other assets before turning to narrative construction. As P3 explained:

> "*I'd probably define all the assets and the scenes. Then once I did that, I would tell Doki the story I want to tell and have it fill in the rest, like the story outline for me.*"

A third group described a more exploratory and freeform approach. For these participants, Doki supported writing and discovery without preparation. P4 mentioned that he

can "*just begin writing something.*" Similarly, P2 worked in small pieces and refined step by step:

> "*I worked on small pieces iteratively, building scenes and refining character interactions step by step.*"

Across these practices, Doki enabled participants to begin video authoring in ways that matched their own practices. Whether starting from a script, from defined assets, or from improvisational writing, participants found that the system supported flexible entry points into the creative process.

4.6.3.2 High reliance on AI

Eight participants reported that they would begin by using AI to generate a first textual draft before refining further. Six participants went further, relying almost entirely on Doki's AI agents throughout the process. They described starting with a rough, often one-line idea, using the AI as a jump start, and then remaining in the conversation panel or inline agent to complete their projects with minimal manual edits. P8 noted that she often preferred asking the AI to revise drafts rather than editing them manually. As she explained, "*AI knows better how to talk to AI.*" P10 described a similar reliance on delegation: "*I don't really read the script the agent creates. I just want to see how the image would look...I generate those images and then see if the agent misunderstood my description.*"

This pattern was not limited to casual creators. P2, a professional filmmaker who produced some of the best videos in our study, revealed that nearly the entire process was conducted with the AI: "*"I had just one line idea...and it really helped me to brainstorm much more in terms of how different scenes will be, how we can create different shots...and then we worked together to turn [it] into a full script.*"

Four participants described a gradual shift from manual work to heavier AI reliance as they grew more comfortable and found it reliable and easy to use. P4 explained:

> "*The workflows are slightly different. At first, I did that very manually. The second one was more hybrid type, and then I'll move to AI more and more, for the last video, I'll just type the command, tell AI what to do rather than manually edit.*"

We interpret this increasing reliance not only as a reflection of participants' trust in the system, but also as a side effect of the design of Doki itself. By making collaboration with AI extremely simple through a text representation, participants can find it natural to delegate more and more of the process to AI. In other words, simplicity of interaction could encourage reliance.

4.6.3.3 Users rely almost entirely on AI still feel ownership.

An interesting finding is that, using Doki, even participants who relied almost completely on AI still felt *strong sense of ownership* of their videos. P4 cited Andy Warhol's saying:

> "*Art is the process of an artist selecting.*"

P1 similarly explained: "*Because I get to decide what every item looks like, I feel like this is my creation. That's what gives me ownership.*" Participants often compared themselves to directors, as P6 described that she feels her role was no different from a director's, except working with an AI agent instead of human agents.

Several participants also contrasted this experience with other platforms. P10 noted, "*When I use MidJourney, if other people use similar prompts, they would get 90% similar images. I would think this is generated by MidJourney, so I give the credit to MidJourney. With Doki, I feel ownership. Every choice of image or cut made me feel I'm a video maker.*" Participants found that Doki enabled a workflow where creators could delegate most production to AI while still experiencing a strong sense of authorship.

Two participants did not use AI at all. Both were animators who typically worked

by hand and had little prior experience with text-to-image or text-to-video models. They expressed a preference for full creative control: "*I just want to create the story by myself.*" (P5).

4.6.4 RQ3. How do users' prior experience with video editing and generative AI influence the ways they perceive and use Doki?

4.6.4.1 Doki empowers participants without previous experience to create video stories

For participants without training in filmmaking, animation, or generative AI, Doki opened up new possibilities that had previously felt out of reach. Four participants noted that they would never have created a video story without Doki, and now, they could express ideas in new ways. For example, P10, who had no prior drawing or animation experience, described:

> "*Like just use this and generate some animation. I do not have experience drawing those, but I can use this tool to create that.*"

Others emphasized Doki's role in unlocking creativity and enabling personal expression. P9 called it "*my dream tool*," explaining that it helped deliver messages and bring childhood stories to life. For many participants, Doki meant more than just a tool – it represented a shift in what creative expression could mean for them. As P1 reflected:

> "*It kind of [offers] a new capability I didn't have before. In the past, if I want to share something with my friends, I have to write it down, but now I can create a video to express whatever I want.*"

4.6.4.2 Experts position Doki as complementary rather than substitutive

Participants with professional experience in video editing described Doki as fundamentally different from their existing tools. Rather than viewing it as a replacement, they see it as

complementary for projects outside their primary workflows. P1 explained,

> "*I think creating a video with Doki is very different than my previous video editing experience. I would even define them as two completely different tasks, and are for very different purposes.*"

Others emphasized this difference by situating Doki within the production pipeline. P4 viewed traditional editing tools to be essential for precision and polish, while Doki was valued for its speed in ideation and draft generation. Similarly P5 stated:

> "*I think I will use [Doki in] the first period of my animation making process because I can first have a script and then input it into Doki to create a video for me to as a reference.*"

Yet even as professionals stressed that Doki could not substitute for high-fidelity outputs, they also highlighted the dramatic contrast in efficiency. P5, who normally works through labor-intensive frame-by-frame animation, reflected that producing one minute of video by hand could take two months, while with Doki she could create it in about an hour. For her, the speed advantage was undeniable, even though the resulting output still carried a "*clearly AI feeling.*"

Another recurring challenge for professionals was that their existing expertise felt difficult to transfer into Doki. P10 reported, "*Although I learned some filming technique, I cannot use it here. Even if I have some idea in my mind, I don't know how to describe in words.*" Similarly, P7 expressed frustration that drawing skills could not be applied directly. For professional with high quality standards, the visual quality Doki can generate still falls short:

> "Animation industry standard is way higher for now...for animators, this is far from enough." – P7

4.6.4.3 Prior experience with generative tools leads to greater satisfaction

Participants with previous experience in generative AI video tools reported greater appreciation for Doki's capabilities. P1, who had recently worked on a generative video research project, explained:

> "*I tried a lot of different tools, including Google's Veo 3, Google Flow, and a tool that we developed, [but] the consistency that I can generate across the scenes is just uncanny. Doki did something that was not previously possible.*"

Similarly, P2 noted that long-term familiarity with generative tools gave him a clearer sense of their strengths and limitations, a perspective that new users may lack:

> "*Now that I have been using these tools for more than two years, I still don't have an idea how good they are with each of these styles...new users would have no idea.*"

Film knowledge also emerged as an important factor. Participants with professional or semi-professional backgrounds were more adept at structuring stories and translating them into coherent visual sequences. As P2 explained,

> "*It usually requires a craftsmanship which most people don't have. People do have ideas, but people don't know how to create an entire piece. That's where expertise comes in.*"

4.7 Discussion and Future Work

While text-native authoring made it easier for participants to get from idea to video, we found that a low-barrier to creation does not on its own result in high-quality storytelling.

Approachability does not guarantee quality. A text-native interface lowers barriers, but strong stories still demand craft. In our diary study, participants moved quickly from concepts to completed videos, yet not all videos had compelling arcs. Doki's document-centric structure makes working with narrative elements like characters and scenes explicit and editable. Expert film makers know how to structure a narrative and how to sequence shots and scenes, while novices are left unsure how to improve their video. Future work includes optional narrative scaffolds (*e.g.*three-act templates, A/B/C plots) and lightweight diagnostics (pacing irregularities, under-specified protagonists) that help authors see and improve a story without adding UI complexity.

Do we really need a video model for longer videos? Current text-to-video systems cap out at short durations (5–10 seconds[50, 113]). At first glance, this seems like the primary limitation. Yet the average shot duration in contemporary film is roughly 4 seconds [33], well within today model's capability. Our formative and diary studies suggest that the main gaps lie elsewhere: *consistency* (stable characters, props, styles), *control* (camera grammar, timing, transitions), and *context* (memory across shots and sequences). In Doki, we treat duration as a compositional rather than monolithic model problem. Text generates *keyframes* that anchor identity, layout, and style; these keyframes expand into short, controllable shots. Longer models will still matter for specific cases (e.g., "oners," dance, sports, talking heads), but for many scenarios, progress on cross-shot memory, identity preservation, and story-aware control is likely to deliver more value than raw clip length.

Document as a human–AI common ground. Most current AI creation tools such as Lovable (text-to-website), Deep Research (text-to-report), or Eleven Labs (text-to-audio), follow a direct input–output model: a prompt produces a finished artifact with limited opportunities for inspection or revision along the way. By contrast, Doki introduces the document as an intermediate representation: a format that is simultaneously readable and editable by humans and interpretable and executable by AI.

In future iterations, even if Doki begins from a simple prompt box, the output would not be a locked timeline or rendered video, but rather a document representation: a shared working space between human and AI where operations remain legible, revisable, and easy for user intervention.

Doki's document representation approach highlights a broader opportunity for the HCI community: how to design effective *intermediate spaces* that make AI activity transparent and at the same time create richer opportunities for human agency.

Temporal expressivity limits of a document representation. A linear document is limited in expressing any temporal concurrency, cross-cut constructs, and duration-sensitive rhythm. Participants struggled to specify overlapping or offset audio (e.g., pre-lapped dialogue, music that bridges sequences), and transitions that bind across shot boundaries (e.g., J/L cuts, cross-dissolves, match cuts) without resorting to external tools. We implemented inline audio notations (§4.4.3.6) to address basic needs, but finer control remains awkward. One path forward is to extend the Doki language with lightweight temporal primitives that preserve readability while increasing control: inline event offsets (e.g., `[SFX door slam +1.2s]` or `[MUSIC fade in 4s]`), paragraph- or sequence-scoped beat/tempo markers (e.g., `#Tempo=110 BPM`), and parameterized transition definitions that bind explicitly to adjacent shots (e.g., `#CrossDissolve 12f`, `#JCut`).

Compositional structures vs. single representation Most video authoring tools rely on "compositional structures" with interconnected underlying data. Professional non-linear editors such as Premiere Pro and Final Cut present projects as a network of linked views: an asset library, a multi-track timeline, a source monitor, and often a transcript panel for editing by text. Descript [36] aligns transcript editing with a timeline. Recent research systems adopt the same principle. VideOrigami [23] links scripts, storyboards, canvases, and timelines.

Doki explores a radically different direction. Instead of coordinating across multiple views, it centralizes all representations within a single text-native substrate. Authoring video is reframed as writing. This presents a different set of tradeoffs. Multiple views can indeed support task-specific efficiency, such as precise pacing control in a timeline or spatial layout on a canvas. But Doki, for a lot of people whether they are professionals or amateurs, greatly simplifies the interface and interaction: there are very few mechanisms to learn and people can begin authoring almost immediately. This unified interface also strengthens version control and collaboration between humans or AI, while introducing challenges in temporal limits and etc. From an HCI perspective, these paradigms are not in opposition but explore distinct possibilities. In a sense, they embody two different philosophies: compositional structures seek to optimize every step of authoring with specialized views, while Doki leverages AI to pursue an ultimately minimalistic interface.

4.8 Conclusion

In this paper, we introduced Doki, a text-native interface for authoring generative videos. Findings from a week-long diary study show that it enables faster idea-to-content workflows, improved coherence via parameterization, and clearer narrative comprehension compared to traditional tools.

More broadly, we beleive that Doki contributes a paradigm that positions text not only as input to generative systems but as the primary substrate for narrative, structure, and production. We believe that as generative models continue to expand in capability, it is essential to rethink interface paradigms for content creation. Doki demonstrates one such direction – an low-barrier, text-native model of authoring – and opens new questions about how natural language might serve as the foundation for future creative tools.

CHAPTER 5

Conclusion

This chapter concludes the thesis by returning to its central claim: that turn-taking has become a default and largely unexamined design constraint in contemporary human–AI interaction. Across three systems, this dissertation demonstrates how relaxing this constraint along different dimensions opens new interaction possibilities that better align with the capabilities of modern AI systems and the practices of human collaborators. This final chapter synthesizes the thesis contributions and outlines directions for future research in HCI and human–AI interaction.

5.1 Review of Thesis Contributions

This thesis makes three primary contributions.

First, it introduces a conceptual framing of *turn-taking* as a design constraint rather than an inevitable feature of AI systems. While turn-taking has deep roots in human conversation and early computing interfaces, this work articulates how its underlying assumptions, including single-threaded interaction, AI reactivity, and transient exchanges, shape and limit the behavior of current human–AI systems. By making these assumptions explicit, the thesis provides a lens for critically examining existing interfaces and for systematically exploring alternatives beyond the chat-based paradigm.

Second, the thesis proposes a set of interaction structures that move beyond turn-taking, grounded in theories from communication, linguistics, and Computer-Supported Coopera-

tive Work. These structures emphasize parallel communication channels that allow multiple streams of meaning to coexist, proactive participation driven by internal processes rather than explicit prompts, and persistent shared spaces that accumulate work over time. Together, they articulate a broader design space for human–AI interaction.

Third, the thesis provides empirical evidence for the value and trade-offs of these interaction structures through the design, implementation, and evaluation of three systems: *Visual Captions*, *Inner Thoughts*, and *Doki*. Visual Captions demonstrates how parallel visual channels can augment live conversation without interrupting conversational flow. Inner Thoughts shows how proactive AI participation can be grounded in intrinsic motivation rather than turn-taking prediction. Doki illustrates how persistent, text-native representations can support long-term collaboration and sensemaking beyond transient exchanges. Across controlled studies, simulations, and in-the-wild deployments, these systems surface both benefits and tensions that inform future design of interactive AI systems.

5.2 Future Work

This thesis opens several directions for future research in HCI and human–AI interaction.

Integrating multiple beyond-turn-taking structures within a single system. Each system in this thesis focuses on relaxing one primary constraint of turn-taking. An important next step is to design and study systems that integrate parallel communication, proactive participation, and persistent shared spaces within a unified interface. For example, future work could explore collaborative tools in which AI agents proactively contribute visual, textual, or structural suggestions inside a persistent workspace while coordinating across multiple parallel channels. Such systems would raise new research questions around orchestration, conflict resolution, and user control, and would be well suited for future CHI or UIST investigations of complex, real-world collaboration.

User control, legibility, and trust in proactive AI behavior. As AI systems move beyond reactive turn-taking, issues of transparency, predictability, and user control become increasingly important. Future research could explore interaction techniques that make AI intentions, internal reasoning, or confidence legible without overwhelming users. This includes studying how users negotiate boundaries around interruption, how they recover from inappropriate AI initiative, and how trust develops over extended use. Designing mechanisms for adjustable autonomy and socially appropriate proactivity remains a central challenge for human–AI collaboration.

Extending beyond-turn-taking interaction to new domains and modalities. While this thesis focuses on conversation and creative authoring, the proposed interaction structures are broadly applicable. Future work could investigate beyond-turn-taking designs in domains such as education, data analysis, programming, accessibility, or collaborative sensemaking. In addition, extending these ideas beyond text and speech to multimodal settings involving gesture, gaze, or augmented and mixed reality offers opportunities to rethink how AI participates in human activity across different temporal and spatial scales. Such explorations can further test and refine the generality of the design principles articulated in this thesis.

REFERENCES

[1] PJ Ace. Pj ace youtube channel, 2025.

[2] Acqualia Software OÜ. Soulver – natural-language notepad calculator. `https://soulver.app`, 2025. Accessed: 2025-09-04.

[3] James E Allen, Curry I Guinn, and Eric Horvtz. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications*, 14(5):14–23, 1999.

[4] Salvatore Andolina, Valeria Orso, Hendrik Schneider, Khalil Klouche, Tuukka Ruotsalo, Luciano Gamberini, and Giulio Jacucci. Investigating proactive search support in conversations. In *Proceedings of the 2018 Designing Interactive Systems Conference*, DIS '18, page 1295–1307, New York, NY, USA, 2018. Association for Computing Machinery.

[5] Salvatore Andolina, Valeria Orso, Hendrik Schneider, Khalil Klouche, Tuukka Ruotsalo, Luciano Gamberini, and Giulio Jacucci. Searchbot: Supporting voice conversations with proactive search. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '18, page 9–12, New York, NY, USA, 2018. Association for Computing Machinery.

[6] Michael Argyle. Non-verbal communication in human social interaction. *Non-verbal communication*, 2:1, 1972.

[7] Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, and George Foster. Retranslation versus streaming for simultaneous translation. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 220–227, Online, July 2020. Association for Computational Linguistics.

[8] Aaron Bangor, Philip T Kortum, and James T Miller. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction*, 24(6):574–594, 2008.

[9] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1:71–81, 2009.

[10] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Kyle Lacey, Yam Levi, Li Cheng, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint*, arXiv:2506.15742v2, June 2025.

[11] Daniel E. Berlyne. *Conflict, Arousal, and Curiosity*. McGraw-Hill, 1960.

[12] Keping Bi, Qingyao Ai, and W Bruce Croft. Asking clarifying questions based on negative feedback in conversational search. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 157–166, 2021.

[13] Dan Bohus and Eric Horvitz. Learning to predict engagement with a spoken dialog system in open-world settings. In *Proceedings of the SIGDIAL 2009 Conference*, pages 244–252, 2009.

[14] Dan Bohus and Eric Horvitz. Models for multiparty engagement in open-world dialog. In *Proceedings of the SIGDIAL 2009 conference, the 10th annual meeting of the special interest group on discourse and dialogue*, page 10, 2009.

[15] Dan Bohus and Eric Horvitz. Multiparty turn taking in situated dialog: Study, lessons, and directions. In *Proceedings of the SIGDIAL 2011 Conference*, pages 98–109, 2011.

[16] Simone Borsci, Alessio Malizia, Martin Schmettow, Frank Van Der Velde, Gunay Tariverdiyeva, Divyaa Balaji, and Alan Chamberlain. The chatbot usability scale: the design and pilot of a usability scale for interaction with ai-based conversational agents. *Personal and ubiquitous computing*, 26:95–119, 2022.

[17] Paul T Brady. A statistical analysis of on-off patterns in 16 conversations. *Bell System Technical Journal*, 47(1):73–91, 1968.

[18] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[19] John Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[20] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

[21] Penelope Brown. *Politeness: Some universals in language usage*, volume 4. Cambridge university press, 1987.

[22] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[23] Yining Cao, Yiyi Huang, Anh Truong, Hijung Valentina Shin, and Haijun Xia. Compositional structures as substrates for human-ai co-creation environment: A design approach and a case study. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–25, 2025.

[24] Stuart K Card, Jock D Mackinlay, and George G Robertson. A Morphological Analysis of the Design Space of Input Devices. *ACM Transactions on Information Systems (TOIS)*, 9(2):99–122, 1991.

[25] Paul Chandler and John Sweller. Cognitive load theory and the format of instruction. *Cognition and instruction*, 8(4):293–332, 1991.

[26] Paul Chandler and John Sweller. The split-attention effect as a factor in the design of instruction. *British Journal of Educational Psychology*, 62(2):233–246, 1992.

[27] Shuo-yiin Chang, Bo Li, Tara N Sainath, Chao Zhang, Trevor Strohman, Qiao Liang, and Yanzhang He. Turn-taking prediction for natural conversational speech. *arXiv preprint arXiv:2208.13321*, 2022.

[28] Jiajian Chen, Jun Xiao, and Yuli Gao. ISlideShow: a Content-Aware Slideshow System. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*, IUI '10, pages 293–296, New York, NY, USA, 2010. Association for Computing Machinery.

[29] Peggy Chi, Tao Dong, Christian Frueh, Brian Colonna, Vivek Kwatra, and Irfan Essa. Synthesis-assisted video prototyping from a document. In *UIST '22: Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, 2022.

[30] Peggy Chi, Nathan Frey, Katrina Panovich, and Irfan Essa. Automatic Instructional Video Creation From a Markdown-Formatted Tutorial. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, UIST '21, page 677–690, New York, NY, USA, 2021. Association for Computing Machinery.

[31] Peggy Chi, Zheng Sun, Katrina Panovich, and Irfan Essa. Automatic Video Creation From a Web Page. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, page 279–292, New York, NY, USA, 2020. Association for Computing Machinery.

[32] Pei-Yu Chi and Henry Lieberman. Intelligent Assistance for Conversational Storytelling Using Story Patterns. In *Proceedings of the 16th International Conference on Intelligent User Interfaces*, IUI '11, page 217–226, New York, NY, USA, 2011. Association for Computing Machinery.

[33] James E Cutting. The evolution of pace in popular movies. *Cognitive research: principles and implications*, 1(1):30, 2016.

[34] Maira Gatti de Bayser, Paulo Cavalin, Claudio Pinhanez, and Bianca Zadrozny. Learning multi-party turn-taking models from dialogue logs. *arXiv preprint arXiv:1907.02090*, 2019.

[35] Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. A survey on proactive dialogue systems: Problems, methods, and prospects. *arXiv preprint arXiv:2305.02750*, 2023.

[36] Descript, Inc. Descript. `https://www.descript.com/`, 2025.

[37] Ruofei Du, David Li, and Amitabh Varshney. Geollery: A mixed reality social media platform. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery.

[38] Ruofei Du, David Li, and Amitabh Varshney. Project geollery.com: Reconstructing a live mirrored world with geotagged social media. In *The 24th International Conference on 3D Web Technology*, Web3D '19, page 1–9, New York, NY, USA, 2019. Association for Computing Machinery.

[39] Starkey Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283, 1972.

[40] Starkey Duncan Jr and George Niederehe. On signalling that it's your turn to speak. *Journal of experimental social psychology*, 10(3):234–247, 1974.

[41] Paul Ekman and Wallace V Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *semiotica*, 1(1):49–98, 1969.

[42] Erik Ekstedt and Gabriel Skantze. Turngpt: a transformer-based language model for predicting turn-taking in spoken dialog. *arXiv preprint arXiv:2010.10874*, 2020.

[43] Jonathan St BT Evans. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, 59(1):255–278, 2008.

[44] Cecilia E Ford and Sandra A Thompson. Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. *Studies in interactional sociolinguistics*, 13:134–184, 1996.

[45] Alisa Fortin, Guillaume Vernade, Kat Kampf, and Ammaar Reshi. Introducing gemini 2.5 flash image (aka "nano-banana"). Technical report, Google AI / Vertex AI, August 2025.

[46] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B. Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. Text-based editing of talking-head video. *ACM Transactions on Graphics*, 38(4):68:1–68:14, 2019.

[47] Jianfeng Gao, Michel Galley, and Lihong Li. *Neural approaches to conversational AI: Question answering, task-oriented dialogues and social chatbots.* Now Foundations and Trends, 2019.

[48] Erving Goffman. *Interaction Ritual: Essays on Face-to-Face Behavior*. Doubleday, 1967.

[49] Yihong Gong and Xin Liu. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, page 19–25, New York, NY, USA, 2001. Association for Computing Machinery.

[50] Google DeepMind. Veo 3. Web page, 2025. State-of-the-art text-to-video generation model capable of producing synchronized audio—including dialogue, sound effects, and ambient noise—with high realism and prompt adherence.

[51] H. P. Grice. Logic and conversation. In *Syntax and semantics*, volume 3, pages 41–58. Academic Press, 1975.

[52] Zhiwei Guan, Shirley Lee, Elisabeth Cuddihy, and Judith Ramey. The validity of the stimulated retrospective think-aloud method as measured by eye tracking. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 1253–1262, 2006.

[53] Sandra G Hart and Lowell E Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*, volume 52, pages 139–183. Elsevier, New York, NY, USA, 1988.

[54] Kaveh Hassani and Won-Sook Lee. Visualizing Natural Language Descriptions: a Survey. *ACM Computing Surveys (CSUR)*, 49(1):1–34, 2016.

[55] Charles T Hemphill, John J Godfrey, and George R Doddington. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.

[56] Ilyena Hirskyj-Douglas, Anna Kantosalo, Andrés Monroy-Hernández, Joelle Zimmermann, Michael Nebeling, and Mar Gonzalez-Franco. Social AR: Reimagining and Interrogating the Role of Augmented Reality in Face to Face Social Interactions. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, pages 457–465, New York, NY, USA, 2020. Association for Computing Machinery.

[57] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[58] Jim Hollan and Scott Stornetta. Beyond being there. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 119–125, 1992.

[59] Karen Holtzblatt and Hugh Beyer. *Contextual design: defining customer-centered systems*. Elsevier, 1997.

[60] Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 159–166, New York, NY, USA, 1999. Association for Computing Machinery.

[61] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[62] Mina Huh, Saelyne Yang, Yi-Hao Peng, Xiang'Anthony' Chen, Young-Ho Kim, and Amy Pavel. Avscript: Accessible video editing with audio-visual scripts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2023.

[63] Ink & Switch. Embark – dynamic documents for making plans. `https://www.inkandswitch.com/embark/`, 2025. Accessed: 2025-09-04.

[64] Ink & Switch. Potluck – dynamic documents as personal software. `https://www.inkandswitch.com/potluck/`, 2025. Accessed: 2025-09-04.

[65] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. Analysis of respiration for prediction of" who will be next speaker and when?" in multi-party meetings. In *Proceedings of the 16th international conference on multimodal interaction*, pages 18–25, 2014.

[66] Yu Jiang, Jing Liu, Zechao Li, Changsheng Xu, and Hanqing Lu. Chat With Illustration: a Chat System With Visual Aids. In *Proceedings of the 4th International Conference on Internet Multimedia Computing and Service*, ICIMCS '12, page 96–99, New York, NY, USA, 2012. Association for Computing Machinery.

[67] Zhenhui Jiang and Izak Benbasat. The Effects of Presentation Formats and Task Complexity on Online Consumers' Product Understanding. *Mis Quarterly*, 1:475–500, 2007.

[68] Dan Kieft. Dan kieft youtube channel, 2025.

[69] Donghyun Kim, Youbin Ahn, Wongyu Kim, Chanhee Lee, Kyungchan Lee, Kyong-Ho Lee, Jeonguk Kim, Donghoon Shin, and Yeonsoo Lee. Persona expansion with commonsense knowledge for diverse and consistent response generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1139–1149, 2023.

[70] Kling AI. Kling ai: Ai image & video maker. `https://www.klingai.com/`, 2025.

[71] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. Will You Accept an Imperfect Ai? Exploring Designs for Adjusting End-User Expectations of Ai Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, New York, NY, USA, 2019. Association for Computing Machinery.

[72] Rachna Konigari, Saurabh Ramola, Vijay Vardhan Alluri, and Manish Shrivastava. Topic shift detection for mixed initiative response. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 161–166, 2021.

[73] Vladimir Kulikov, Shahar Yadin, Matan Kleiner, and Tomer Michaeli. Sinddm: A single image denoising diffusion model, 2022.

[74] Kenichi Kumatani, Sankaran Panchapagesan, Minhua Wu, Minjae Kim, Nikko Strom, Gautam Tiwari, and Arindam Mandai. Direct modeling of raw audio with dnns for wake word detection. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 252–257. IEEE, 2017.

[75] John E Laird. *The Soar cognitive architecture*. MIT press, 2019.

[76] Mackenzie Leake, Hijung Valentina Shin, Joy O. Kim, and Maneesh Agrawala. Generating Audio-Visual Slideshows From Text Articles Using Word Concreteness. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–11, New York, NY, USA, 2020. Association for Computing Machinery.

[77] James R. Lewis and Jeff Sauro. The factor structure of the system usability scale. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 53, pages 1259–1263. SAGE Publications, 2009.

[78] Jing Li, Yang Chen, Rui Zhao, et al. Promptcanvas: Composable prompting workspaces using dynamic widgets. *arXiv preprint arXiv:2506.03741*, 2025.

[79] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[80] Jian Liao, Adnan Karim, Shivesh Singh Jadon, Rubaiat Habib Kazi, and Ryo Suzuki. Realitytalk: Real-time speech-driven augmented presentation for ar live storytelling. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, New York, NY, USA, 2022. Association for Computing Machinery.

[81] Lizi Liao, Grace Hui Yang, and Chirag Shah. Proactive conversational agents in the post-chatgpt world. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3452–3455, 2023.

[82] Xingyu "Bruce" Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang "Anthony" Chen, and Ruofei Du. Visual captions: Augmenting verbal communication with on-the-fly visuals. In *Proceedings of the 2023 CHI Conference on Human*

Factors in Computing Systems, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[83] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.

[84] Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. Towards conversational recommendation over multi-type dialogs. *arXiv preprint arXiv:2005.03954*, 2020.

[85] John K Local, John Kelly, and William HG Wells. Towards a phonology of conversation: turn-taking in tyneside english1. *Journal of Linguistics*, 22(2):411–437, 1986.

[86] Luma Labs. Luma ai – 3d capture and generative ai. `https://lumalabs.ai`, 2025.

[87] Kent Lyons, Christopher Skeels, Thad Starner, Cornelis M Snoeck, Benjamin A Wong, and Daniel Ashbrook. Augmenting Conversations Using Dual-Purpose Speech. In *Proceedings of the 17th Annual ACM Symposium on User Interface Software and Technology*, pages 237–246, New York, NY, USA, 2004. Association for Computing Machinery.

[88] Richard E Mayer. Multimedia Learning. In *Psychology of Learning and Motivation*, volume 41, pages 85–139. Elsevier, New York, NY, USA, 2002.

[89] David H McFarland. Respiratory markers of conversational interaction. 2001.

[90] Harry McGurk and John MacDonald. Hearing Lips and Seeing Voices. *Nature*, 264(5588):746–748, 1976.

[91] Christian Meurisch, Cristina A Mihale-Wilson, Adrian Hawlitschek, Florian Giger, Florian Müller, Oliver Hinz, and Max Mühlhäuser. Exploring User Expectations of Proactive AI Systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4):1–22, 2020.

[92] Rada Mihalcea and Chee Wee Leong. Toward Communicating Simple Sentences Using Pictorial Representations. *Machine Translation*, 22(3):153–173, 2008.

[93] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International conference on machine learning*, pages 7721–7735. PMLR, 2021.

[94] Florian Müller, Sebastian Günther, Azita Hosseini Nejad, Niloofar Dezfuli, Mohammadreza Khalilbeigi, and Max Mühlhäuser. Cloudbits: Supporting conversations

through augmented zero-query search visualization. In *Proceedings of the 5th Symposium on Spatial User Interaction*, SUI '17, page 30–38, New York, NY, USA, 2017. Association for Computing Machinery.

[95] Notion Labs Inc. Notion. `https://www.notion.com/product/docs`, 2025.

[96] Hiromu Ogawa and Pattie Maes. Smartwatch-Based Topic Suggestions to Enrich Casual Conversations in Awkward Encounters. In *Proceedings of the 2020 International Symposium on Wearable Computers*, pages 68–72, New York, NY, USA, 2020. Association for Computing Machinery.

[97] OpenAI. Introducing canvas in chatgpt, 2024.

[98] OpenAI. Learning to reason with llms. September 2024.

[99] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

[100] Yi-Hao Peng, Ming-Wei Hsi, Paul Taele, Ting-Yu Lin, Po-En Lai, Leon Hsu, Tzu-chuan Chen, Te-Yen Wu, Yu-An Chen, Hsien-Hui Tang, and Mike Y. Chen. SpeechBubbles: Enhancing Captioning Experiences for Deaf and Hard-of-Hearing People in Group Conversations. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–10, New York, NY, USA, 2018. Association for Computing Machinery.

[101] Christopher Peters. Direction of attention perception for conversation initiation in virtual environments. In *International Workshop on Intelligent Virtual Agents*, pages 215–228. Springer, 2005.

[102] Christopher Peters, Catherine Pelachaud, Elisabetta Bevacqua, Maurizio Mancini, and Isabella Poggi. A model of attention and interest using gaze behavior. In *International Workshop on Intelligent Virtual Agents*, pages 229–240. Springer, 2005.

[103] Pika. Pika labs – ai video generation. `https://pika.art`, 2025.

[104] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, New York, NY, USA, 2021. PMLR, Association for Computing Machinery.

[105] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125:10, 2022.

[106] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084:10, 2019.

[107] Xuhui Ren, Hongzhi Yin, Tong Chen, Hao Wang, Zi Huang, and Kai Zheng. Learning to ask appropriate questions in conversational recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 808–817, 2021.

[108] Bradley Rhodes and Thad Starner. Remembrance Agent: a Continuously Running Automated Information Retrieval System. In *The Proceedings of the First International Conference on the Practical Application of Intelligent Agents and Multi Agent Technology*, volume 1, pages 487–495, New York, NY, USA, 1996. Association for Computing Machinery.

[109] Frank E Ritter, Farnaz Tehranchi, and Jacob D Oury. Act-r: A cognitive architecture for modeling cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 10(3):e1488, 2019.

[110] Sonia Roccas, Lilach Sagiv, Shalom H Schwartz, and Ariel Knafo. The big five personality factors and personal values. *Personality and social psychology bulletin*, 28(6):789–801, 2002.

[111] Steve Rubin, Floraine Berthouzoz, Gautham J Mysore, Wilmot Li, and Maneesh Agrawala. Content-based tools for editing audio stories. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pages 113–122, 2013.

[112] Jurgen Ruesch, Gregory Bateson, Eve C Pinsker, and Gene Combs. *Communication: The social matrix of psychiatry*. Routledge, 2017.

[113] Runway. Runway – ai magic tools for creators. `https://runwayml.com`, 2025.

[114] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *language*, 50(4):696–735, 1974.

[115] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

[116] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *arXiv preprint arXiv:2104.07636*, 1, 2021.

[117] Nazmus Saquib, Rubaiat Habib Kazi, Li-Yi Wei, and Wilmot Li. Interactive Body-Driven Graphics for Augmented Video Performance. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, New York, NY, USA, 2019. Association for Computing Machinery.

[118] AI Video School. Ai video school youtube channel, 2025.

[119] Samira Shaikh, Tomek Strzalkowski, George Aaron Broadwell, Jennifer Stromer-Galley, Sarah M Taylor, and Nick Webb. Mpc: A multi-party chat corpus for modeling social phenomena in discourse. In *LREC*. Citeseer, 2010.

[120] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[121] Gabriel Skantze. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language*, 67:101178, 2021.

[122] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

[123] Salvador Soto-Faraco, Alan Kingstone, and Charles Spence. Multisensory Contributions to the Perception of Motion. *Neuropsychologia*, 41(13):1847–1862, 2003.

[124] Barry E Stein and M Alex Meredith. *The Merging of the Senses*. The MIT press, New York, NY, USA, 1993.

[125] Kevin Stratvert. Kevin stratvert youtube channel, 2025.

[126] Lucille Alice Suchman. *Plans and situated actions: The problem of human-machine communication*. Cambridge university press, 1987.

[127] William H Sumby and Irwin Pollack. Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, 26(2):212–215, 1954.

[128] S Shyam Sundar. Multimedia Effects on Processing and Perception of Online News: a Study of Picture, Audio, and Video Downloads. *Journalism & Mass Communication Quarterly*, 77(3):480–499, 2000.

[129] Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric P Xing, and Zhiting Hu. Target-guided open-domain conversation. *arXiv preprint arXiv:1905.11553*, 2019.

[130] Louis Ten Bosch, Nelleke Oostdijk, and Lou Boves. On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, 47(1-2):80–86, 2005.

[131] Bekzat Tilekbay, Saelyne Yang, Michal Lewkowicz, Alex Suryapranata, and Juho Kim. Expressedit: Video editing with natural language and sketching. arXiv:2403.17693 (ACM IUI 2024), 2024.

[132] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[133] David Traum, Antonio Roque, Anton Leuski, Panayiotis Georgiou, Jillian Gerten, Bilyana Martinovski, Shrikanth Narayanan, Susan Robinson, and Ashish Vaswani. Hassan: A virtual human for tactical questioning. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 71–74, 2007.

[134] Anh Truong, Floraine Berthouzoz, Wilmot Li, and Maneesh Agrawala. Quickcut: An interactive tool for editing narrated video. In *UIST '16: Proceedings of the 29th Annual ACM Symposium on User Interface Software and Technology*, pages 497–507, 2016.

[135] Bret Victor. Explorable explanations. In *Proceedings of the ACM Conference on User Interface Design*, 2006.

[136] Marilyn Walker and Steve Whittaker. Mixed initiative in dialogue: An investigation into discourse segmentation. *arXiv preprint cmp-lg/9504007*, 1995.

[137] Hanna M. Wallach. Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 977–984, New York, NY, USA, 2006. Association for Computing Machinery.

[138] Bryan Wang, Yuliang Li, Zhaoyang Lv, Haijun Xia, Yan Xu, and Raj Sodhi. Lave: Llm-powered agent assistance and language augmentation for video editing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 699–714, 2024.

[139] Miao Wang, Guo-Wei Yang, Shi-Min Hu, Shing-Tung Yau, and Ariel Shamir. Write-a-video: Computational video montage from themed text. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 38(6), 2019.

[140] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–14, 2021.

[141] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[142] Jimmy Wei, Kurt Shuster, Arthur Szlam, Jason Weston, Jack Urbanek, and Mojtaba Komeili. Multi-party chat: Conversational agents in group settings with humans and models. *arXiv preprint arXiv:2304.13835*, 2023.

[143] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45, jan 1966.

[144] Morton Wiener and Albert Mehrabian. *Language Within Language: Immediacy, a Channel in Verbal Communication*. Ardent Media, New York, NY, USA, 1968.

[145] Allison Woodruff and Paul M Aoki. How push-to-talk makes talk less pushy. In *Proceedings of the 2003 ACM International Conference on Supporting Group Work*, pages 170–179, 2003.

[146] Haijun Xia. Crosspower: Bridging graphics and linguistics. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 722–734, New York, NY, USA, 2020. Association for Computing Machinery.

[147] Haijun Xia, Jennifer Jacobs, and Maneesh Agrawala. Crosscast: Adding Visuals to Audio Travel Podcasts. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, page 735–746, New York, NY, USA, 2020. Association for Computing Machinery.

[148] Haijun Xia, Jennifer Jacobs, and Maneesh Agrawala. Crosscast: Adding visuals to audio travel podcasts. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST '20)*, 2020. Short paper.

[149] Jun Xu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. Conversational graph grounded policy learning for open-domain conversation generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1835–1845, 2020.

[150] Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. Realpersonachat: A realistic persona chat corpus with interlocutors' own personalities. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 852–861, 2023.

[151] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[152] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

[153] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation, 2022.

[154] Jezia Zakraoui, Moutaz Saleh, and Jihad Al Ja'am. Text-to-picture tools, systems, and approaches: a survey. *Multimedia Tools and Applications*, 78:22833–22859, 2019.

[155] Saizheng Zhang. Personalizing dialogue agents: I have a dog, do you have pets too. *arXiv preprint arXiv:1801.07243*, 2018.

[156] Xiaojin Zhu, Andrew B Goldberg, Mohamed Eldawy, Charles R Dyer, and Bradley Strock. A Text-to-Picture Synthesis System for Augmenting Communication. In *AAAI*, volume 7, pages 1590–1595, New York, NY, USA, 2007. Association for Computing Machinery.

[157] Douglas E. Zongker and David H. Salesin. On Creating Animated Presentations. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '03, pages 298–308, Goslar, DEU, 2003. Eurographics Association.

[158] Changqing Zou, Haoran Mo, Chengying Gao, Ruofei Du, and Hongbo Fu. Language-Based Colorization of Scene Sketches. *ACM Transactions on Graphics*, 38(6):233:1–233:16, Dec. 2019.