Democratizing Content Creation and Consumption through Human-AI Copilot Systems

Bryan Wang
University of Toronto
Toronto, Canada
bryanw@dgp.toronto.edu

ABSTRACT

Content creation and consumption play vital roles in our lives. However, creating high-quality content can be challenging for beginners, while navigating through and consuming vast amounts of media content can be overwhelming and cumbersome. My Ph.D. research focuses on democratizing content creation and improving content consumption experiences for everyday users. I achieve this by designing and evaluating interactive AI systems that serve as copilots, assisting users with tedious tasks. I explore various media modalities, such as video, audio, text, and images, and investigate how their interplay can address problems in individual modalities. This paper offers a comprehensive overview of my research agenda, including recent contributions, on-going progress, and future directions.

KEYWORDS

Content Creation and Consumption, Human-AI Copilot System, Creativity Support Tools.

ACM Reference Format:
Bryan Wang. 2023. Democratizing Content Creation and Consumption through Human-AI Copilot Systems. In *The 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23 Adjunct), October 29-November 1, 2023, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3586182.3616707

1 INTRODUCTION

Digital content pervades our daily routines, from morning social media scrolling to learning recipes from YouTube, consuming podcasts during commutes, and sharing mealtime photos. The ability to produce and consume content has become an integral part of our lives, enabling us to connect, learn, and express ourselves. Despite the ease of capturing and sharing content through mobile devices and social media, creating quality content remains challenging, requiring skills in capturing, editing, storytelling, etc. Moreover, as the digital content landscape expands, efficiently consuming relevant content has become increasingly important. For example, quickly seeking and consuming information of interest in podcasts or videos spanning hours.

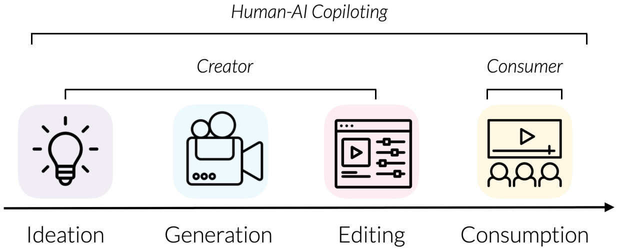


Figure 1: My research focuses on creating and studying human-AI co-pilot systems applicable to challenges within the entire lifecycle of digital content, from ideation to generation, editing, and consumption.

In my Ph.D. research, I develop human-AI copilot systems to address practical challenges in creating and consuming diverse content. Using my systems, humans and AI work together as copilots, with AI automating tasks and providing suggestions while humans refine and correct AI output, ultimately achieving what neither can accomplish alone. My long-term goal is to make content creation and consumption easier for all, particularly novices without specialized training, to democratize these processes. My research scope encompasses the entire life cycle of creative content. Figure 1 illustrates several prominent stages my research has covered: 1) ideation and planning, 2) generation and capture of raw content, 3) editing process, which transforms these raw materials into polished, engaging media, and 4) consuming content in a personalized and efficient manner.

I work with various media modalities, including video, audio, imagery, and text. These modalities encompass diverse content forms, from instructional videos and audio narratives to visual arts. Notably, there tends to exist an intrinsic interconnection among different modalities, presenting possibilities for integrating cross-modal AI. Much of my work harnesses AI within one modality to address problems in the interconnected modalities.

Recent AI breakthroughs, particularly in generative AI, are transforming the landscape of media content. Through this research, I aim to explore the emerging opportunities and challenges they bring to content creation and consumption, thereby contributing to the extensive lineage of HCI research on creativity tools and, more broadly—Human-AI interaction.

UIST '23 Adjunct, October 29-November 1, 2023, San Francisco, CA, USA

ACM ISBN 979-8-4007-0096-5/23/10.
https://doi.org/10.1145/3586182.3616707

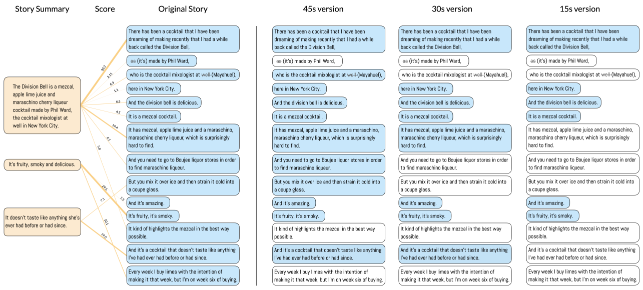
2 PAST CONTRIBUTIONS

2.1 Content Creation

I have been leading and involved in research in AI-assisted content creation [1, 3, 6]. My research emphasizes the collaborative nature of humans and AI throughout the creative process. In addition, my work increasingly incorporates large language models (LLMs), enabling users not only to receive assistance from AI but also to guide the AI toward specific outcomes through language [1, 3]. In the following sections, I discuss three systems: ROPE [6], Promptify [1], and Stargazer [3], which exemplify my research approaches to the area. Each of them addresses different aspects of the creative workflow for various types of content.

2.1.1 Simplifying Short-form Audio Story Editing. Digital content on emerging social media like Instagram Reels and TikTok has become shorter and more concise. However, creating concise content poses challenges, especially for new creators who struggle to edit their material to fit different platform restrictions. In Record Once, Post Everywhere (ROPE) [6], I address the challenges creators face when authoring short-form audio stories by developing ROPE, an interactive AI system that assists users in condensing voice recordings to a specific target length while maintaining high audio quality. The key idea behind the system is a novel approach that formulates audio shortening as a combinatorial optimization problem, aiming to select optimal sentence combinations that adhere to length constraints (Figure 2). ROPE transcribes the audio, applies neural abstractive summarization to extract key topics, and assigns scores based on sentence importance. These scores, combined with sentence durations, guide the optimization algorithm to select the most suitable sentences. ROPE also allows users to refine the algorithm's output through an interactive process. They can modify automatic suggestions directly or specify new constraints for the system to re-run the optimization process. Study results show that ROPE can generate high-quality edits, alleviating the cognitive loads of creators for shortening content.

2.1.2 Facilitating Text-to-Image Generation. Text-to-image generative models such as Stable Diffusion [4] can produce high-quality images based on natural language descriptions. However, crafting prompts–the primary means of steering image generation–remains a challenging task. Promptify [1] addresses the challenges associated with prompt writing by allowing the user to work alongside an LLM to ideate possible subjects and styles in the generated images. By few-shot prompting an LLM, i.e., providing the model with a few target task examples to guide its generation, Promptify takes inspiration from prompts shared in the online community to assist users in expanding their prompts with effective keywords. The generated images are visualized on an interface that allows users to organize and browse them flexibly (Figure 3). Promptify also suggests keywords based on the generated images, helping users refine their original prompts to achieve desired features and avoid undesired ones. Altogether, the system incorporates a feedback loop that enables users to iteratively refine their prompts and enhance specific features. Promptify assists in three creative stages in text-to-image generation workflow from ideation, generation, to editing in a unified system.



The Division Bell is a mezcal, apple lime juice and maraschino cherry liqueur cocktail made by Phil Ward, the cocktail mixologist at well in New York City.

It's fruity, smoky and delicious.

It doesn't taste like anything she's ever had before or had since.

There has been a cocktail that I have been dreaming of making recently that I had a while back called the Division Bell,

~~as~~ (it's) made by Phil Ward,

who is the cocktail mixologist at ~~well~~ (Mayahuel),

here in New York City.

And the division bell is delicious.

It is a mezcal cocktail.

It has mezcal, apple lime juice and a maraschino, maraschino cherry liqueur, which is surprisingly hard to find.

And you need to go to Boujee liquor stores in order to find maraschino liqueur.

But you mix it over ice and then strain it cold into a coupe glass.

And it's amazing.

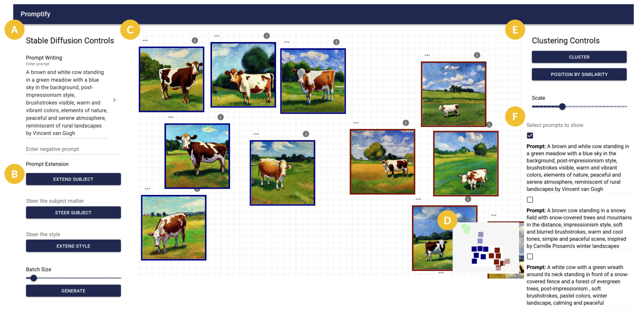
It's fruity, it's smoky.

It kind of highlights the mezcal in the best way possible.

And it's a cocktail that doesn't taste like anything I've had ever had before or had since.

Every week I buy limes with the intention of making it that week, but I'm on week six of buying.

Figure 2: Users record audio stories once, and ROPE [6] will automatically shorten them to create multiple shorter versions. Users can also easily refine the automatic suggestions.



A brown and white cow standing in a green meadow with a blue sky in the background, post-impressionism style, brushstrokes visible, warm and vibrant colors, elements of nature, peaceful and serene atmosphere, reminiscent of rural landscapes by Vincent van Gogh

☑ **Prompt:** A brown and white cow standing in a green meadow with a blue sky in the background, post-impressionism style, brushstrokes visible, warm and vibrant colors, elements of nature, peaceful and serene atmosphere, reminiscent of rural landscapes by Vincent van Gogh

☐ **Prompt:** A brown cow standing in a snowy field with snow-covered trees and mountains in the distance, impressionism style, soft and blurred brushstrokes, warm and cool tones, simple and peaceful scene, inspired by Camille Pissarro's winter landscapes

☐ **Prompt:** A white cow with a green wreath around its neck standing in front of a snow-covered fence and a forest of evergreen trees, post-impressionism , soft brushstrokes, pastel colors, winter landscape, calming and peaceful

Figure 3: Promptify [1] leverages LLMs to assist users in iteratively exploring and refining text-to-image generation prompts.

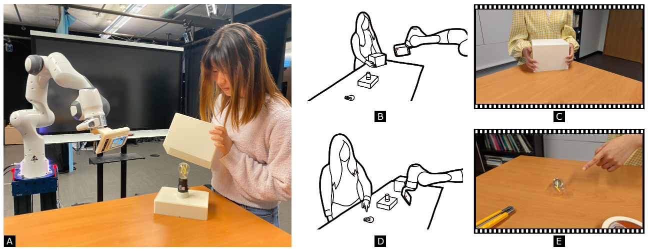


Figure 4: Stargazer [3] uses a camera robot to automatically track areas of interest during recording, enabling instructors to seamlessly incorporate camera control commands through gestures and speech, enhancing the creation of dynamic shots.

2.1.3 Assisting Real-Time Video Capture. Video tutorials are an effective means of teaching physical skills. However, to effectively convey the subtleties of the skill being taught, it is often necessary to film demonstrations from multiple perspectives. This presents a challenge for instructors who do not have access to a dedicated cameraperson, as they must work within the constraints of stationary cameras. To tackle this issue, Stargazer [3] introduces a robotic camera system that assists in recording instructional videos for tabletop tasks (Figure 4). With Stargazer, instructors can seamlessly guide camera operations in real time, ensuring a smooth flow of instruction. For example, when an instructor mentions, *"If you take a closer look, you’ll notice this socket has a hexagon shape,"* the Stargazer system automatically performs a zoom-in camera operation. This capability is achieved by few-shot prompting an LLM to interpret the instructor’s narrations and automatically adjust camera angles and zooming during the presentation. Stargazer effectively addresses challenges in the content generation stage, where the raw video footage is being produced.

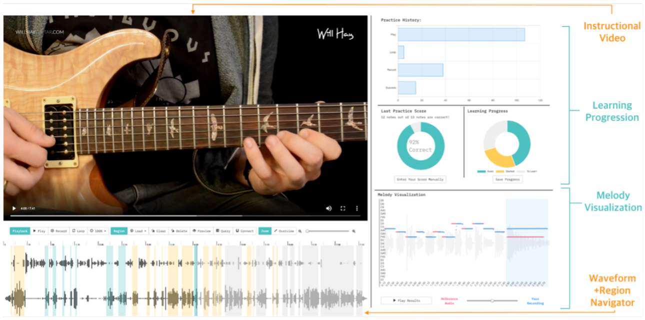


Figure 5: Soloist [9] extracts musical information from raw videos, providing interactive visualizations for video navigation and real-time performance feedback to enhance the user’s learning experience.

2.2 Content Consumption

With the growing volume of digital content available on the internet, efficient content consumption is increasingly crucial. My past research highlights two vital aspects of consuming content: navigation and comprehension. Navigation involves efficiently finding necessary information, while comprehension involves understanding and converting information into personal knowledge or skills, particularly for learning and educational purposes.

2.2.1 Navigating and Learning with Music Instructional Videos. Instructional videos are a prevalent content format for learning musical instruments, but practicing alongside these videos can pose challenges for users. Common video navigation techniques are not specifically designed for the context of practicing instruments with videos. Additionally, the absence of immediate feedback and personalization in pre-recorded videos can obstruct learning. To tackle these issues, I designed Soloist [9], an automatic pipeline that employs deep-learning-based audio processing to extract musical information from raw guitar instructional videos, thereby generating interactive tutorials. Soloist segment videos into sections containing demonstrations and offers a suite of efficient video navigation techniques designed to expedite music practice. Furthermore, Soloist’s interface records user performance and gives immediate feedback by comparing user performance with tutor demonstrations (Figure 5). Similar to ROPE [6], Soloist supports mixed-initiative interactions [2] where users can correct any potential errors made by the AI. Our user study suggested that participants unanimously preferred learning with Soloist over learning with raw videos.

2.2.2 Efficient GUI Information Seeking and Consumption. A thread of my research focuses on leveraging AI to enable people to interact with graphical user interfaces (GUIs) through natural language. On the surface, GUIs may not appear to fit into the conventional categories of content. However, they can be viewed as a compilation of organized content comprising various common elements such as text, images, and videos. Despite their widespread use as a dominant user interface, many GUIs rely primarily on visual means to convey information. As a result, information consumption can become problematic when visual channels are unavailable or when users face an overwhelming amount of information.

My work has addressed the challenges by bridging GUIs with natural language. In Screen2Words [8], I propose a novel task called screen summarization, which generates succinct language descriptions of mobile UIs. Screen summaries can support conversational agent applications and augment screen readers. To realize this, I collected a dataset of 112k language summaries across 22k UI screens and trained multimodal deep neural networks to generate screen summaries. The study results showed our methods surpassing heuristic baselines, and multimodal approaches outperforming unimodal ones. Our dataset and model code are open-sourced for future research.

In a subsequent study, I investigated the potential of pre-trained LLMs as an additional interface between humans and GUIs, allowing versatile information exchange through natural language [7]. I developed prompting techniques to adapt an LLM to mobile UIs and tested them in various conversational UI tasks. Our results highlighted how LLMs can accelerate UI information consumption. For example, they can accurately summarize the screen and answer users’ questions regarding specific information on the screen. This thread of work provides users with alternative ways to consume and interact with information presented on GUIs and can potentially contribute to GUI accessibility research.

3 CURRENT AND FUTURE WORK

Moving forward, I outline several areas that I am currently exploring and will continue to work on in the future.

3.1 Emerging Content Formats and Devices

Creative content is constantly evolving, driven by shifts in content formats and capturing devices. The rising popularity of short-form content, which engages viewers in seconds, serves as a notable example. Oftentimes, the emergence of new content formats also introduces fresh authoring challenges. ROPE [6] exemplifies my work on tackling these new challenges within the domain of short-form audio story creation. Building upon this foundation, my future work will continue addressing the authoring and consumption difficulties associated with emerging content formats. The emergence of innovative capturing devices, such as the camera-mounted robotic arm used by Stargazer [3] for tutorial videos, and on-glasses cameras on wearables for ego-centric capture, expand content creation possibilities. Additionally, augmented and virtual reality headsets offer potential for 3D content creation. I am actively exploring these fresh avenues for content creation.

3.2 Paradigm Shifts in Content Editing

Over recent decades, content editing paradigms have transformed significantly, with creators now less burdened by handling raw materials, which often require higher cognitive loads. Video and audio editing used to involve managing individual clips or waveforms in timeline-based editors, which can be labor-intensive. Thanks to HCI research, more intuitive paradigms have emerged. For example, text-based editing allows users to edit video and audio by modifying a time-aligned transcript. This approach offers a semantically intuitive format and has been successfully integrated into commercial products. In my on-going research, I ask: What will be the next paradigm shift in content editing? A potential answer is agent-based, or co-pilot, editing paradigms, which alleviate the need for users to edit content frame by frame or word by word. Instead, they simply specify higher-level editing directions and let AI handle the actual edits. ROPE [6] is a step in this direction, offering an automatic algorithm that provides editing suggestions and allows easy user refinement of automated output. With the advent of generative models that are controllable via language, we are seeing more content creation tools that generate impressive results based on high-level user commands. I plan to delve further and contribute to the next paradigm shift in content editing. Interestingly, this potential shift could revisit the enduring debate between direct manipulation and interface agents [5].

3.3 Bridging Content Modalities

As demonstrated in my previous work and extensive research in the field, the interconnectedness of various content modalities offers significant value in addressing challenges within each modality. Videos consist of visual and audio elements, while audio can include speech and music. Speech can also be converted into text. These connections open up innovative approaches to tackling issues in each modality. For example, Soloist [9] segments guitar instructional videos by analyzing their soundtracks. ROPE [6] suggests edits to audio based on its language content. Promptify [1] assists users in generating images using text prompts. As the community focuses on foundation models that integrate multiple modalities into a single framework, the potential for multimodal solutions in content creation and consumption will continue to grow.

3.4 Personalized Content Consuming Experience

Media content is often consumed in a one-size-fits-all manner. Once crafted by its creator, the content remains static and is consumed identically by everyone, irrespective of their diverse backgrounds. Current technology allows for some customization in content consumption, such as speed alterations or manual skipping, but these options often prove unsatisfactory and cumbersome. Soloist [9] is a step towards customized consumption and learning with prerecorded videos using interactive AI systems. I will continue exploring techniques that enable consumers to tailor their consumption according to their preferences or consuming habits. For instance, when listening to a podcast, a system could automatically bypass sections containing familiar information. While reading an article, a system could extract and reorganize pertinent information based on the reader's previous reading patterns.

3.5 Broader Implications of AI in Content Creation and Consumption

As we integrate AI more pervasively into content creation and consumption, it is vital to consider its broader implications beyond functional capabilities and address potential ramifications such as user autonomy, attribution for creations, and ethical concerns. There are research questions that I am keen to study. For example, how do generative AIs and automated creation tools, such as one-click solutions, influence users' sense of agency and ownership over the content they generate across different contexts? Additionally, when generative AI is employed for content creation, how should credit allocation be approached? Should contributors to model training data receive credits, and if so, how should this be implemented? Although my past research has not explored these areas, I recognize their significance and plan to explore these themes in future research.

4 CONCLUSION

In this paper, I have outlined my Ph.D. research on democratizing content creation and improving content consumption experiences through the development of human-AI copilot systems. I have highlighted my past contributions and discussed my ongoing and future plans in my long-term research agenda.

REFERENCES

[1] Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-Image Generation through Interactive Prompt Exploration with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3586183.3606725

[2] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. https://doi.org/10.1145/302979.303030

[3] Jiannan Li, Maurício Sousa, Karthik Mahadevan, Bryan Wang, Paula Akemi Aoyagui, Nicole Yu, Angela Yang, Ravin Balakrishnan, Anthony Tang, and Tovi Grossman. 2023. Stargazer: An Interactive Camera Robot for Capturing How-To Videos Based on Subtle Instructor Cues. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 800, 16 pages. https://doi.org/10.1145/3544548.3580896

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

[5] Ben Shneiderman and Pattie Maes. 1997. Direct Manipulation vs. Interface Agents. *Interactions* 4, 6 (nov 1997), 42–61. https://doi.org/10.1145/267505.267514

[6] Bryan Wang, Zeyu Jin, and Gautham Mysore. 2022. Record Once, Post Everywhere: Automatic Shortening of Audio Stories for Social Media. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 14, 11 pages. https://doi.org/10.1145/3526113.3545680

[7] Bryan Wang, Gang Li, and Yang Li. 2023. Enabling Conversational Interaction with Mobile UI Using Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 432, 17 pages. https://doi.org/10.1145/3544548.3580895

[8] Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. 2021. Screen2Words: Automatic Mobile UI Summarization with Multimodal Learning. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 498–510. https://doi.org/10.1145/3472749.3474765

[9] Bryan Wang, Meng Yu Yang, and Tovi Grossman. 2021. Soloist: Generating Mixed-Initiative Tutorials from Existing Guitar Instructional Videos Through Audio Processing. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 98, 14 pages. https://doi.org/10.1145/3411764.3445162