

ID.8: Co-Creating Visual Stories with Generative AI

VICTOR NIKHIL ANTONY, Johns Hopkins University, U.S.A

CHIEN-MING HUANG, Johns Hopkins University, U.S.A

ID.8: An Integrated Authoring System for Visual Story Creation Aided by Generative AI

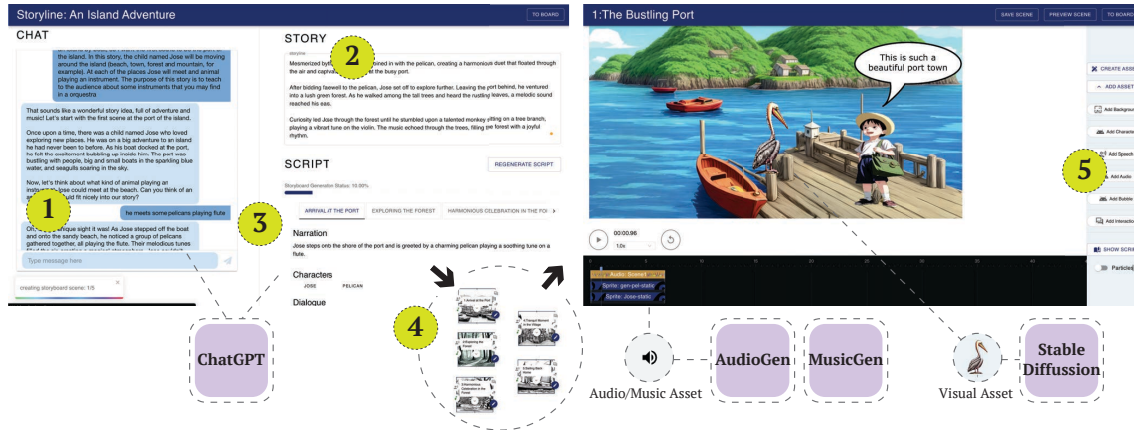


Fig. 1. ID.8 features a multi-stage visual story authoring workflow facilitated by generative AI. (1) The story creation begins with users collaborating with ChatGPT to create a storyline and (2) manually editing of story content for finer edits. (3) Then, ID.8 automatically parses the story co-created with ChatGPT into a scene-by-scene script to be edited further by the user. (4) The story scenes from the script are automatically pre-populated and organized in the Storyboard and (5) edited in the Scene Editor where users use StableDiffusion, AudioGen, and MusicGen to generate story elements and synchronize story element on the canvas and the timeline.

Storytelling is an integral part of human culture and significantly impacts cognitive and socio-emotional development and connection. Despite the importance of interactive visual storytelling, the process of creating such content requires specialized skills and is labor-intensive. This paper introduces ID.8, an open-source system designed for the co-creation of visual stories with generative AI. We focus on enabling an inclusive storytelling experience by simplifying the content creation process and allowing for customization. Our user evaluation confirms a generally positive user experience in domains such as enjoyment and exploration, while highlighting areas for improvement, particularly in immersiveness, alignment, and partnership between the user and the AI system. Overall, our findings indicate promising possibilities for empowering people to create visual stories with generative AI. This work contributes a novel content authoring system, ID.8, and insights into the challenges and potential of using generative AI for multimedia content creation.

CCS Concepts: • **Human-centered computing** → **Interaction design; Interactive systems and tools.**

Authors' addresses: Victor Nikhil Antony, Johns Hopkins University, Baltimore, Maryland, U.S.A; Chien-Ming Huang, Johns Hopkins University, Baltimore, Maryland, U.S.A.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Association for Computing Machinery.

Manuscript submitted to ACM

Manuscript submitted to ACM

Additional Key Words and Phrases: Storytelling, Generative AI, Creativity

1 INTRODUCTION

Storytelling is a defining aspect of the human experience that has been practiced through diverse media, such as written text, oral traditions, cave paintings, and more [51]. Visual stories are narratives that are augmented by various forms of media such as drawings, illustrations, animations, and videos, serving to enhance the overall storytelling experience [9]. Visual stories tend to increase interest in and emotional engagement with the narrative even as they improve understanding and retention of the story’s content [35]; they can be an ideal medium for psycho-educational interventions [5], health communications [56], language learning [50], intergenerational bonding [31], and creative self-expression [48].

Despite the benefits and use cases of visual stories, their creation process remains a challenging, multifaceted task that unfolds via a sequence of essential steps—such as brainstorming to cultivate ideas, scripting to develop the narrative, storyboarding for visual planning, amassing the necessary media assets, piecing together the elements, and refining the content through editing—culminating in the distribution of the finalized story [43]. The nature of multimodal asset creation and the technical demands of specialized software (e.g., Adobe Creative Studio, Final Cut Pro) present a skill barrier that hampers both expert and novice creators from fully tapping into visual stories’ creative potential. Lowering barriers to visual story authoring can enable the production of individualized and customized content that may lead to improved outcomes in varying use cases and amongst diverse populations.

Recent advances in generative AI have enabled the production of text [7], images [37], audio [14], and videos [32] from user instructions; generative AI models hold the potential to help democratize the visual story authoring landscape. Human-AI co-creation is a paradigm wherein human users collaborate with AI with varying degrees of partnership to create a product [45]. Several co-creative systems have explored how generative models may help users create storylines [13, 40], draw visuals [28, 66] and compose music [34]. However, to the best of our knowledge, no systems exist, co-creative or otherwise, that enable the end-to-end authoring of visual stories. Leveraging generative AI to simplify the authoring process may help enable the quick and expressive creation of visual stories.

Toward empowering users to effectively explore the creative possibilities of and quickly iterate over and generate visual stories, we built **ID.8** (*ideate*), an end-to-end authoring system for visual story creation with generative AI integrated into its workflow (see Fig. 1). Our system enables users to collaborate with ChatGPT (a large language model) [7] to co-write a script for the story, generate visual assets via Stable Diffusion (a text-to-image model) [47], and generate audio assets with AudioGen (a text-to-audio model) [27] and MusicGen (a text-to-music model) [14]. ID.8 embodies a “human-in-control, AI-in-the-loop” design framework to balance user autonomy and AI assistance. Our goal is to harmonize the control, agency, content safety, and human touch inherent to the manual story creation process with the creative variability and production efficiency of generative AI.

We conducted a two-phased evaluation of the ID.8 system to assess its usability and creative breadth. We found that ID.8 provided an enjoyable user experience and that users greatly appreciated the value of integrating generative AI in the visual story authoring workflow; moreover, users generated a wide variety of stories via our system, demonstrating its creative capabilities. Through this evaluation, we also gained a deeper understanding of the challenges faced by users while using a multimodal co-creative system, offering insights and design implications for future human-AI co-creative systems. This work makes three key contributions:

- (1) We design, develop, and release ID.8: a novel, open-source, end-to-end system that enables visual story authoring via a unified interface and a human-AI co-creative workflow, aiming to lower the skill ceiling required for and enable the agile iteration of and broader creative expression in visual storytelling.
- (2) Insights from two user evaluations highlight the current opportunities and challenges of multimodal content creation via state-of-the-art generative AI.
- (3) We put forward a set of design guidelines for human-AI co-creative systems based on our experience and empirical evidence from evaluating ID.8.

2 RELATED WORK

Here, we explore the three areas of work related to the motivation and development of ID.8. Section 2.1 highlights how conventional story authoring tools facilitate users in crafting visual stories but often place the burden of asset production on users. Section 2.2 focuses on state-of-the-art generative AI models, exploring how advancements in this area hold the potential for democratizing content creation across modalities. Section 2.3 discusses the potential of human-AI co-creative systems, which synergize the complementary capabilities of humans and AI in a collaborative creative process. Through ID.8, we seek to address the gap at the intersection of these areas by building a platform that leverages generative AI to enhance the visual story authoring experience in a co-creative setting and to help better understand the challenges of co-creation in a multi-modal setting.

2.1 Story Authoring Tools

Visual stories take many forms such as comics, animated videos, interactive informatics, etc. There have been attempts at simplifying the authoring process towards enabling expressive storytelling. PrivacyToons [52] and DataToon [24] allow users to create comics by incorporating their own sketches and adjust the arrangement and panels through pen and touch interactions; PrivacyToons moreover supports the design and creative ideation yet remains limited in terms of viewer interaction. However, these systems only allow for creation of static stories. ScrollyVis [36] is a web app for authoring, editing, and presenting data-driven scientific articles and allows for the integration of common sources such as images, text, and video; it supports multiple layers, dynamic narratives, however, is limited to creating interactive articles. Katika [22] is an end-to-end system that simplifies the process of creating explainer motion graphics videos for amateurs; it provides a graphical user interface that allows users to create shots based on a script, add artworks and animation from a crowdsourced library, and edit the video using semi-automated transitions. These systems demonstrate the productivity improvements that story authoring tools provide to users however are still limited due to the burden of asset production being placed on the user.

To bridge the gap that exists due to burden of assets generation being placed on end-users and limited interactivity, we designed ID.8 and sought to leverage generative AI to support each stage of the visual story authoring process to empower users to quickly explore the creative landscape and materialize their vision.

2.2 Generative AI and Content Generation

Breakthroughs in generative AI capabilities have opened up a plethora of possibilities across various domains specifically in supporting creative processes and democratizing content generation. Advancements in text generation capabilities using large language models (LLMs) (e.g., GPT-3 [7], PaLM [12], LLaMA [55], etc) have been transformative for natural language processing tasks ranging from text summarization to code generation [19]; the versatility in executing a range of vastly different natural language tasks such as summarization and question-answering makes these LLMs

particularly useful [30]. In the context of storytelling, LLMs are beginning to be used for generating narrative structures, dialogues, and even full-length screenplays [1, 49]. While they should not replace human creativity entirely, they offer new avenues for brainstorming, initial drafting, and expediting the scriptwriting process [26].

In the visual domain, text-to-image models (*e.g.*, Stable Diffusion [47], Midjourney, Dall-E [44]) have been employed to create photo-realistic paintings, animated illustrations, and even designs for fabrics or interiors [62, 67]. Moreover, image generation models are capable of accepting inputs beyond just text—for instance, sketches, images, masks, canny edges, pose points—to help guide the generation of images or to edit existing work thus potentially enabling expressive communication of intent from users [58, 68]. Such technology can allow for rapid prototyping and experimentation, granting creatives, both amateurs and professionals, a new range of tools to express their artistic visions.

Generative models have also been used for composing music that spans various genres, including classical, jazz, and electronic music and are even capable of generating more abstract audio effects such as battlefield backdrops. AI systems like OpenAI’s Jukebox [15] and recent models such as Meta’s MusicGen [14] and AudioGen [27] are designed to either generate new compositions or assist composers in their creative processes.

Despite the opportunities these state-of-the-art generative AI models represent, there is a lack of work on the incorporation of generative AI that supports content generation in various modalities into a unified authoring process and environment for visual story creation. With ID.8, we aim to integrate the growing capabilities of generative models to empower individuals to create interactive multimodal content to support visual storytelling.

2.3 Human-AI Co-Creative Systems

Human-AI co-creation refers to shared, collaborative creation of a creative product by human(s) and AI system(s) [45]. Advancements in generative AI are opening new frontiers for human-AI collaboration leading to a wide array of work towards building and evaluating human-AI co-creative systems in a variety of domains, including music, design, dance, and writing.

Several co-creative systems have explored various methods of supporting human-AI collaborative writing [13, 21, 46, 54]. WordCraft [21] is a text editor with a built-in AI writing assistant which showed the need for goal alignment and adaptability to users’ expertise in co-creative systems. Talebrush [13] was designed to help users intuitively guide the narrative for co-created stories using a line sketching mechanism spotlighting the value of exploring more intuitive mechanisms for facilitating intent communication between the user and the AI. BunCho was built to support high-level and creative writing of stories for a text-based interactive game and was found to enhance the enjoyability of the writing process and to broaden the creative breadth of the stories [40].

Co-creative systems have also been built to support visual artistic expression. Reframer [28] presented a novel human-AI drawing interface that facilitates an iterative, reflective and concurrent co-creation process while SmartPaint [53] introduced a co-creative painting system that empowers amateur artists to turn rough sketches into complete paintings. Virtual embodied conversational agents were incorporated into the co-creative drawing process in Creative PenPal [46] and found to potentially improve user engagement and the collaborative experience. Furthermore, co-creative AI systems have been built to enable creative drawing with children; StoryDrawer is a system that supports collaborative drawing and establishes how co-creative AI can help enhance creative outcomes and lead to higher engagement in the creative process [66]. AIStory, another co-creative system for children, simplified the prompting process for text-to-image models by replacing text input with icons which children could choose to generate visuals to accompany their narratives [20].

Table 1. System Design Criteria (DC) and Rationale for ID.8

<p>Design Criteria 1: Enable End-to-End Authoring</p> <p>Authoring visual stories is a complex, multi-stage process that includes brainstorming, scripting, storyboarding, gathering media assets, synchronizing individual elements, refining the narrative, and delivering the completed story [43]. Given the lack of comprehensive tools in this space, ID.8 consolidates these disparate steps into a unified workflow, thereby empowering users to create visual stories more effortlessly.</p>
<p>Design Criteria 2: Facilitate Intuitive Narrative Assembly</p> <p>The complexity in visual story authoring arises from juggling diverse elements like narrative arcs, and spatio-temporal asset pacing, making it challenging to maintain a coherent view of the interactions of different narrative elements. ID.8 addresses this by helping users establish robust mental models of the narrative structure through clear interaction cues and visual markers among various story elements.</p>
<p>Design Criteria 3: Amplify Creative Exploration</p> <p>Generative models hold the potential for rapid exploration of a vast creative landscape, thereby amplifying individual creativity and imagination [61]. ID.8 is designed to streamline the exploration and curation of these AI-generated artifacts, allowing users to seamlessly integrate these outputs into their creative vision.</p>
<p>Design Criteria 4: Safeguard Creative Autonomy</p> <p>Critical issues like the loss of user autonomy and the generation of potentially dangerous or unsafe creative outputs can manifest in co-creative systems, often due to AI decisions taking precedence over human input [8]. ID.8 is designed to prioritize user control, empowering them to evaluate, select, and integrate AI-generated assets, thus effectively mitigating the risk of AI-driven decisions overriding user choices [61].</p>
<p>Design Criteria 5: Support Modularity and Extensions</p> <p>With the rapid advancements in Generative AI technologies in mind, ID.8 is architected at the system design level for modularity and extensibility, allowing for the seamless integration of emerging state-of-the-art generative models.</p>

Beyond the text and visual domains, prior work has explored the human-AI co-creation in the music domain. COSMIC [69] explored collaborative music generation with a chatbot whereas MoMusic [4] presented a prototype collaborative Music Composition system that uses human motion as a key input signal yet again highlighting the unique domains and interaction modalities co-creative AI can enable.

Together, these human-AI co-creative systems represent the nascence of a type of human-computer interaction that will become increasingly commonplace as the capabilities of generative models continue to expand and shape human creative work, highlighting the importance of designing effective human-AI interaction strategies to harness the power of generative AI in human creative processes. With ID.8, we seek to deepen our understanding of how people interact with multiple generative AI models (particularly LLMs, text-to-image and text-to-audio models) and incorporate their outputs in the creation of visual stories.

3 ID.8: AN INTEGRATED AUTHORING SYSTEM FOR VISUAL STORY CREATION

Design Philosophy and Criteria. ID.8 aims to balance user autonomy and AI assistance through a design philosophy centered on a “human-in-control, AI-in-the-loop” framework. We strive to combine the control, agency, safety and

human touch intrinsic to manual story creation with the creative accessibility and efficiency gains offered by generative AI. To align this philosophy with our goal of simplifying visual story creation, we outline the our core design criteria (DC) in Table 1.

3.1 System Overview

ID.8 features a multi-stage visual story authoring workflow facilitated by generative AI (see Fig. 1). Users initiate story creation with the *Storyline Creator* module where they collaborate with ChatGPT (alias 'Leela') to create the story plot and manually perform finer edits of the co-crafted plot (See Fig. 2-1). Then, ID.8 systematically converts the storyline into a scene-by-scene script to be edited further by the user. The story scenes from the script are automatically pre-populated and organized as individual nodes in the *Storyboard* module. Users then edit each scene in the *Scene Editor* where users use StableDiffusion, AudioGen and MusicGen to generate story elements and synchronize various narrative elements on the canvas and the timeline. Users can watch the story as a whole or preview a single scene to experience the story as the viewer and adjust accordingly.

We developed ID.8 as an open-source¹ ReactJS application to facilitate inter-platform accessibility. ID.8 uses REST API calls to remote servers (*i.e.*, OpenAI API², HuggingFace Spaces³, StableDiffusionWeb API⁴) for interactions with the generative models, providing a flexible architecture that supports seamless model updates as new models being introduced.

3.2 Storyline Creator

The first stage of the ID.8 workflow is the Storyline Creator, a module that serves as a co-screenwriting environment. Within this module, ChatGPT powered chatbot—nicknamed 'Leela'—is pre-configured with an initial prompt to jointly craft narratives with the user (see Fig. 2-1). In this setting, Leela poses questions, offers creative suggestions, and generates narrative components based on user input and feedback. This design enables users to focus on higher-level creative decisions, such as overarching plot design, character development, and thematic elements, while Leela supplements these efforts by producing nuanced narrative details and structural elements (DC1,2). Users retain complete creative control of the co-created narrative (see Fig. 2-2) and are able to directly edit and refine the story generated with Leela (DC3). After the narrative is finalized, ID.8 compiles the co-crafted story into an editable, structured scene-by-scene script, complete with scene titles, narration, character lists, and dialogues—all generated via GPT-3.5. Moreover, ID.8 auto-generates corresponding scene nodes in the Storyboard, each featuring the scene's title and a monochromatic background visualized using Stable Diffusion (see Fig. 2-3). This synthesis aims to bridge the narrative content with the visual scenes, thereby providing users with a foundation for further story construction (DC1).

3.3 Storyboard

As a visual canvas (see Fig. 3), the Storyboard module, supports an interactive, node-based visualization of the narrative (DC2). Users create and link individual nodes, each representing a specific scene, to create a cohesive story structure and flow. Metadata associated with each scene—such as titles, background settings, interactive components, and multimedia elements—are compactly displayed, aiding quick comprehension and navigation (DC2). The interface is designed to

¹We provide access to ID.8 through Docker containers for the review process and will release the system after paper acceptance. Link to google drive folder with instructions and container: <https://tinyurl.com/y3hjswhc>

²<https://openai.com/product>

³<https://huggingface.co/facebook/audiogen-medium>

⁴<https://stablediffusionapi.com/>

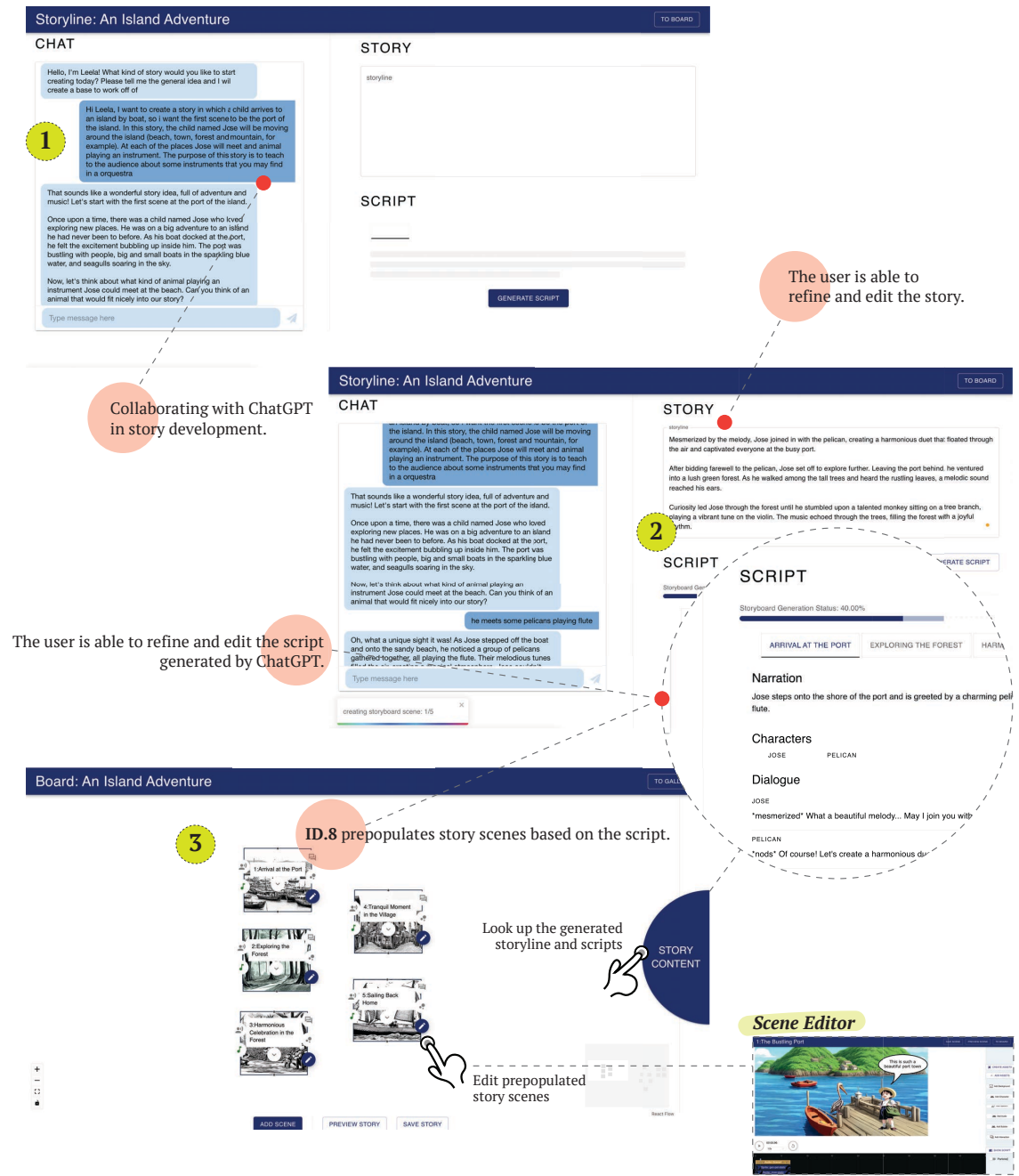


Fig. 2. ID.8 enables generation of a story (1) by collaborating with ChatGPT, and also allows the user (2) to manually edit the story and then (3) generates—using ChatGPT—a structured script and pre-populates the storyboard with scenes from the script.

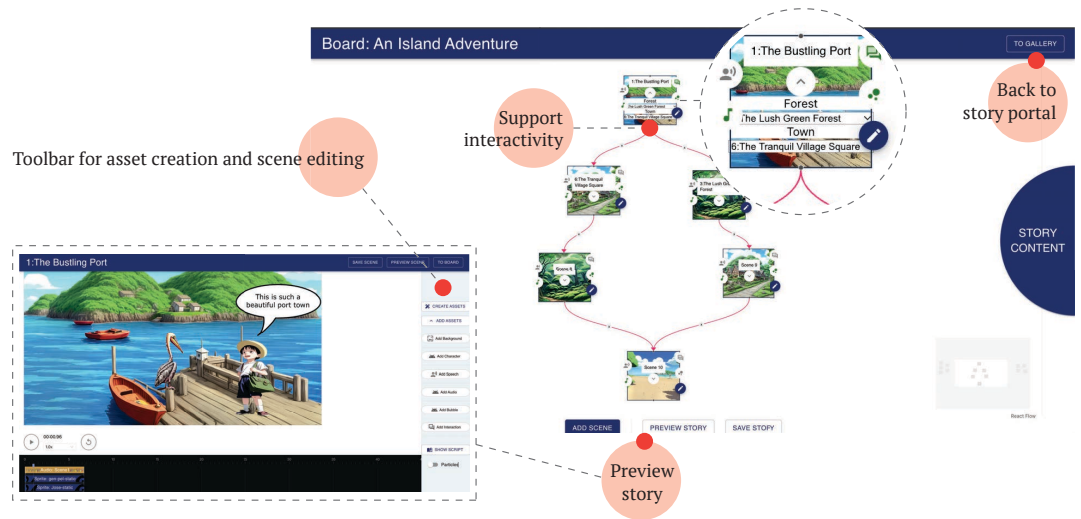


Fig. 3. ID.8 Storyboard allows for organization of the story flow by linking scenes and specifying how story viewer inputs should impact the flow of the story. Users access the Scene Editor module by double-clicking a scene node. Users can also preview their story.

be intuitive, providing features like scene addition, deletion, and replication, as well as specifying a narrative starting point. To enrich the storytelling experience, the Storyboard module supports conditional narrative branching through its interaction components. This functionality aims not only to streamline the user’s creative workflow but also to provide granular control over the narrative trajectory, ensuring that the Storyboard functions as both a planning tool and an interactive blueprint for the story (DC4).

3.4 Scene Editor

The Scene Editor is a multi-faceted workspace that complements both the Storyline Creator and the Storyboard by enabling authoring and editing capabilities for individual scenes. Designed to enable quick creative exploration and iteration, it provides an array of tools to select, synchronize, and experiment with the audio-visual elements and the flow of a scene (DC1,2) (see Fig. 4).

Canvas: The Visual Workspace. At the center of the Scene Editor is the Canvas, serving as a visual workspace for real-time scene assembly. It provides an interactive interface where users can directly view and manipulate visual elements like characters and speech bubbles; an overlaid contextual menu appears when an element is selected, offering a range of customization options (e.g., dimensions, start and end locations for an animation path). Dynamic background effects, mimicking natural phenomena like rain or snow, can also be set using the toolbar (see Fig. 3) to enrich the visual storytelling experience.

Timeline: Orchestrating Timing and Sequencing. Situated below the canvas, the timeline offers a visual platform for coordinating the sequence and timing of scene components (e.g., characters) (DC2). It aids in crafting a cohesive narrative flow by allowing users to chronologically synchronize and quickly adjust these elements, ensuring that the visual and auditory assets sync well with each other and the script.

Asset Creator. Anchored within the Scene Editor, the Asset Creator (see Fig. 4) serves as a platform for generating, selecting, and adapting both visual and audio assets. This tool, featuring a dual-tabbed interface comprising a ‘Visuals

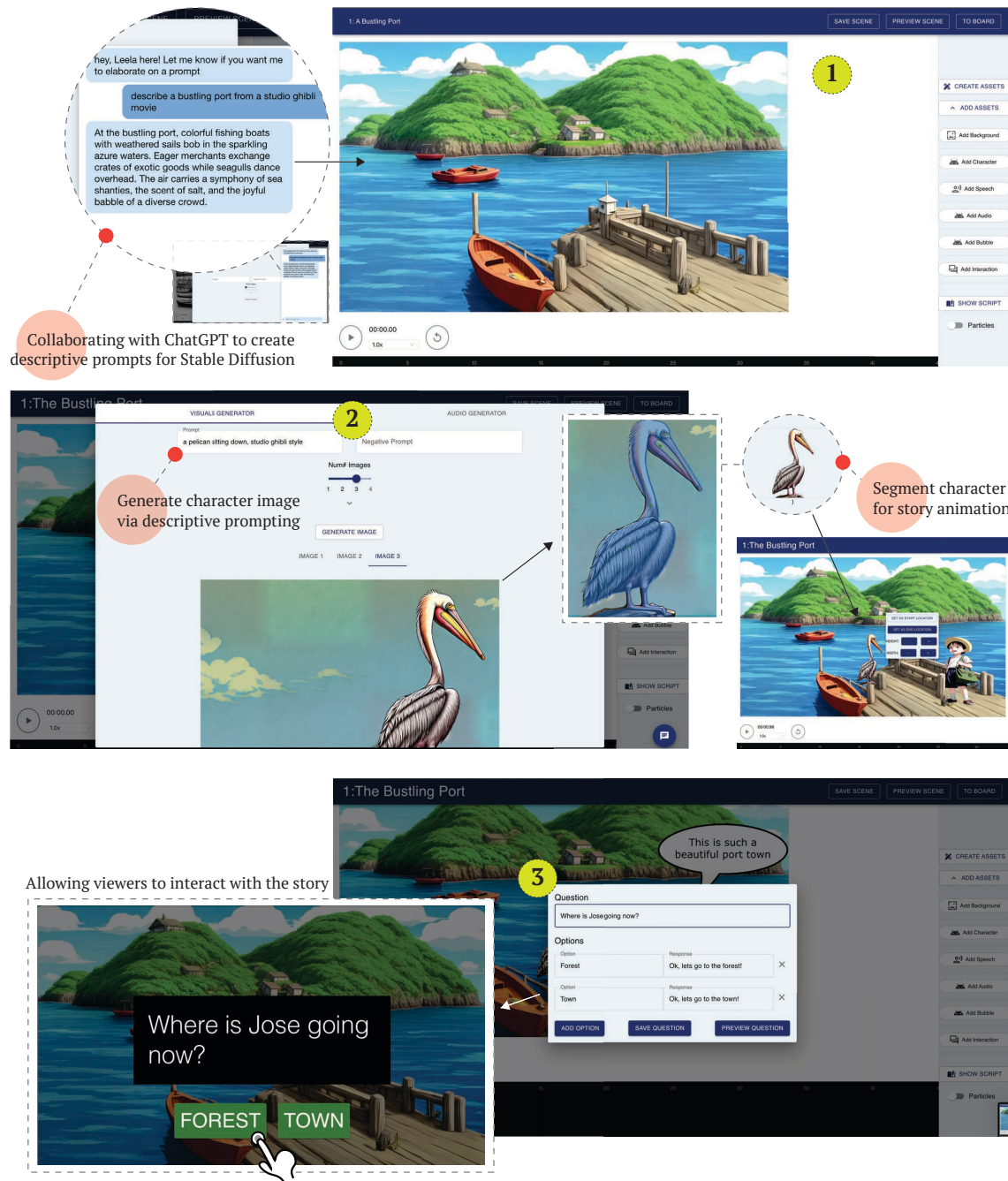


Fig. 4. (1)The ID.8 Scene Editor enables creation of prompts for text-to-image/audio models in collaboration with ChatGPT; (2) For character generation, ID.8 empowers users to select parts of the generated output to be used in the story; (3) ID.8 provides a simple interface for adding interaction with viewer.

Generator’ and an ‘Audio Generator’, allows users to explore creative possibilities and materialize their storytelling vision.

- **Visuals Generator.** Using Stable Diffusion [47], a text-to-image model, the Visuals Generator supports image creation based on user-provided prompts. Users may save generated image(s) as background or extract parts to be saved as characters using Meta’s ‘Segment Anything’ model [25] (see Fig. 4-2). Advanced controls for fine-tuning the model’s output include negative prompts, a variable range for the number of generated images (1–4), denoising steps, and various modes like panorama and self-attention.
- **Audio Generator.** Focused on sound crafting, the Audio Generator is powered by Meta’s AudioGen [27] and MusicGen [14] models, allowing for the creation of sound effects (e.g., applause, wolf howls) and musical pieces (e.g., lo-fi background tracks, classical carnatic compositions) based on user descriptions. Controls like duration (1–10 seconds), top-p and guidance scale afford more granular manipulation of audio generation.

The Asset Creator also integrates ‘Leela’ (ChatGPT) to facilitate collaborative prompt authoring with the user (see Fig. 4-1) towards more effective visual and audio generation thus striving to help users realize their creative vision. Moreover, it offers users preview functionalities and multiple output options for greater creative control (DC3). This setup ensures that the AI components play a supportive yet non-dominant role in the creative process (DC4).

Speech Generator. To enable narration and dialogue, Scene Editor includes a Speech Generation module powered by Google Cloud Text-to-Speech⁵. This feature allows users to select, customize and preview voice outputs as well as adjust parameters such as pitch and speed for creating more tailored voices. Users have the option to save their customized speech profiles to ensure consistent auditory experiences across multiple scenes (DC1).

Viewer Interaction. ID.8 aims to elevate viewer engagement by incorporating interactive elements that invite viewers’ participation within the story (DC1). Towards accomplishing this, the Scene Editor provides an option to append interactive questions at the conclusion of individual scenes, aimed at capturing viewer input. Questions are both displayed on-screen and vocalized through Google Cloud Text-to-Speech, with a set of selectable responses also presented to the viewer. Depending on the selected response, specific auditory feedback is triggered, and the storyline may diverge accordingly. For scenes featuring interactive components, ID.8 enables the creation of conditional narrative branches. Users have the flexibility to dictate how viewer responses can influence subsequent scenes, thereby introducing a level of interactivity that has the potential to impact the story’s direction (DC4). ID.8 provides a stand-alone story viewer platform that is able to play the generated stories. This viewer platform can be used by embodied agents (e.g., virtual characters, social robots) to enrich the story viewing experience; for instance, a social robot can “tell” the story along with expressive movements to engage the viewer in domains such as education and therapy.

4 EVALUATION

We conducted two studies to evaluate ID.8 and to understand how end-users interact and collaborate with generative AI to create visual stories. Study 1 sought to understand the usability of ID.8, the creative support by generative AI in the visual story authoring workflow, and whether users can effectively generate different story elements such as plot, background, characters, audio effects, etc. Study 2 aimed to gain a deeper understanding of the ID.8 creative breath and the co-creative user experience through an open-ended story generation task that participants engage in over a longer period time (i.e., one week) outside of a controlled lab environment. We report details of the two study designs and

⁵<https://cloud.google.com/text-to-speech/>

study-specific findings in this section and describe the general lessons learned from the studies and design guidelines for human-AI co-creative systems in Section 5.

4.1 Study Measures

In both studies, we used two questionnaires to assess the usability and collaborative aspects of ID.8: the System Usability Scale (SUS) [6] and the Mixed-Initiative Creativity Support Index (MICSI) [29]. The SUS is a 10-item, five-point Likert scale questionnaire specifically designed to provide a reliable metric for system usability. SUS Scores for positively worded items are calculated as $(\text{response} - 1)$ and for negatively worded items as $(5 - \text{response})$; these scores are then summed and multiplied by 2.5 to yield a final SUS score ranging from 0 to 100. A SUS score above 70 is considered to signal good usability of a system [3].

MICSI is an 18-item scale created to assess human-machine creative collaboration. It includes five sub-scales for *Creativity Support*—namely Enjoyment, Exploration, Expressiveness, Immersion, and Results-Worth-Effort; each of these sub-scales is evaluated via a pair of seven-point Likert-scale questions, with the score derived from the mean response value for each question pair. Additionally, MICSI includes sub-scales for assessing *Human-Machine Collaboration*, each evaluated through a separate seven-point Likert-scale question; these sub-scales cover Communication, Alignment, Agency, and Partnership. A score of 5 or above on these MICSI sub-scales signifies a positive user experience [29]. It is worth noting that the interpretation of the Agency sub-scale is different from the rest of sub-scales; lower scores suggest the system is perceived primarily as a tool, while higher scores imply it is seen more as a collaborative partner [29]. We also slightly adjusted the accompanying set of exploratory questions in MICSI, which focus on Contribution, Satisfaction, Surprise, and Novelty in relation to the final outcome; specifically, we replaced the term ‘sketch’ with ‘story’ to better align with the context of ID.8.

4.2 Study 1: Evaluating Usability

We conducted an in-lab system evaluation to understand the usability of ID.8, the creative support provided by generative AI in the visual story authoring workflow, and whether users can generate different story elements using ID.8. This evaluation was centered around a structured task that was designed to ensure that users interacted with various sub-components of ID.8.

4.2.1 Study Procedure and Task. After providing informed consent, participants filled out a pre-study questionnaire collecting demographics information (see Section 4.2.2). Then, an experimenter presented a walk-through video demonstrating the system’s functionalities to familiarize participants with ID.8. Following this, participants were allocated 75 minutes to compose a story of their choice, and to visualize three scenes from that story. Participants were instructed to utilize the Asset Creator module to generate at least one background visual, one piece of audio or music, and one static character; they were also asked to add interactive elements to their story. Besides these basic task requirements, participants were free to be as creative as they wanted. To better capture the co-creative process, we gathered interaction logs with the generative models. We also asked participants to keep notes on their experience and creation process during the session. After their allotted time, participants were asked to fill out a post-study questionnaire, which included the SUS and MICSI. Finally, a semi-structured interview was administered to explore in greater depth the participants’ preferences, criticisms, and suggestions for ID.8 in the co-creative process. Participants were compensated with gift cards valued at the rate of \$15/hr.

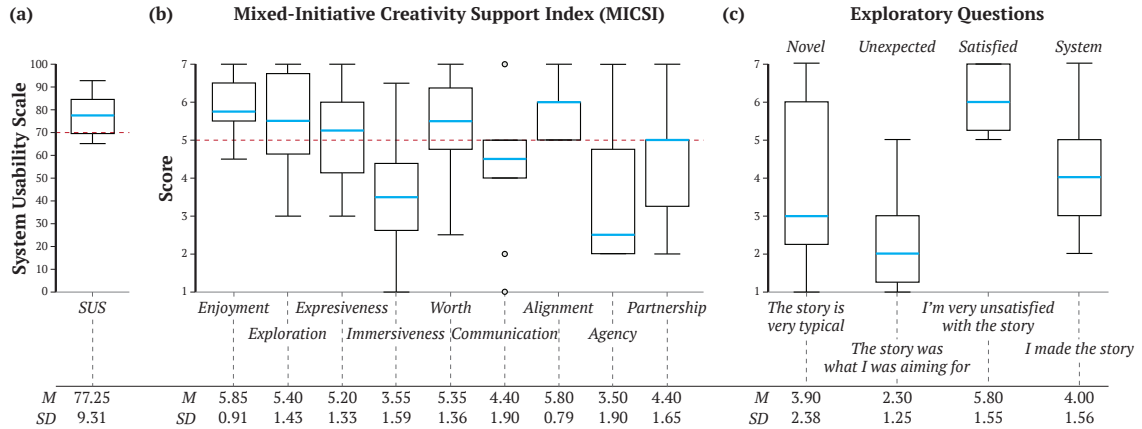


Fig. 5. Results from Study 1: (a) SUS Scores, (b) MICSI Sub-Scale Scores, (c) Exploratory Question Responses.

4.2.2 Participants. We recruited 11 participants (6 female, 5 male) via online chat forums. Participants were aged 19 to 28 ($M = 24.1, SD = 2.10$) and had a wide range of backgrounds and expertise including robotics, law and innovation consultancy. Participants had a varied degree of familiarity with LLMs (very familiar=1, familiar=5, neutral=2, unfamiliar=2), familiarity with image/audio generation models (familiar=1, neutral=4, unfamiliar=1, very unfamiliar=4), and familiarity with creating visual content (very familiar=1, familiar=2, neutral=4, unfamiliar=2, very unfamiliar=1). We excluded one participant from analysis due to a data collection failure during their session.

4.2.3 Results. We provide the quantitative results in Fig. 5 and report on key observations from the analysis of the survey data and transcriptions of the user interviews below.

Participants found ID.8 to be easy to use (SUS score: $M = 77.25, SD = 9.31$) and recognized the value of the system in fast-tracking the visual story authoring. Moreover, participants were able to generate a wide range of scenes (see Fig. 6) with different artistic styles, characters and auditory elements using ID.8 and were satisfied with the authored story as indicated by the responses to the exploratory question regarding story satisfaction (see Fig. 5-c) further highlighting the potential of generative AI in supporting the multi-modal creative process.

P1: “I think the image and the audio generation is quite interesting because yeah, if I would have done it myself, it would take a long time but using it is quite convenient for me and the chat with the LLM is quite useful. And I think the general user interface was easy to use.”

P8: “It was really helpful for quickly iterating on new ideas and exploring potential broad strokes of the story and for thinking about different ways you could represent a character or different events that could happen or the scenery or music. Relative to me trying to sketch those out myself, it was really efficient.”

The MICSI Enjoyment sub-scale scores ($M = 5.85, SD = 0.91$) and the feedback from the interviews show how ID.8 facilitated an enjoyable authoring experience.

P4: “I feel like it’s really cool. It’s a really cool tool. Like I feel like I would use it again, if I had more time to play around with it because it’s really fun.”

The overall positive user experience suggested by the Enjoyment, Exploration, Expressiveness, Worth, and Alignment sub-scales (Fig. 5-b) indicates that ID.8 has reasonable realization of our design objectives of streamlining visual story authoring (DC1), enabling creative exploration (DC3), and maintaining user control (DC4).



Fig. 6. Scenes from stories generated by participants using ID.8 in Study 1 and Study 2.

Participant responses to the exploratory question of who (System or I) made the story, suggest that users perceived their interaction with the system to be a balanced creative endeavor in terms of contribution. Yet, the Partnership score ($M = 4.40, SD = 1.65$) indicates a need for enhancing the sense of collaboration; this is further supported by the low Agency score ($M = 3.50, SD = 1.90$) indicating that the system was generally perceived not as a collaborative partner but rather as a tool. Similarly, the scores for Immersiveness ($M = 3.55, SD = 1.59$) and Communication ($M = 4.40, SD = 1.90$) lagged behind, signaling potential areas for improvement in the co-creative experience design. These lower scores suggest the need to refine the co-creative process to facilitate a more immersive user experience and to enhance mechanisms for effectively capturing user intentions.

4.3 Study 2: Evaluating Co-Creative Potential

In order to better understand the creative width of ID.8 and to evaluate its human-AI co-creative experience in a less controlled setting, we conducted a longer-term, in-the-field study with participants who did not take part in Study 1 where we deployed our system on their personal devices and asked them to create a story over a week using ID.8.

4.3.1 Study Procedure and Task. After providing informed consent, participants filled out a pre-study questionnaire collecting demographics information (see Section 4.2.2). An experimenter deployed ID.8 using Docker images on participants' personal devices. Then, an experimenter presented the walk-through video shown in Study 1, demonstrating the system's functionalities to familiarize participants with ID.8. Participants were asked to use ID.8 to create a story of their choice and visualize it. We collected logs of their interactions with the generative models and asked them to keep a journal of their process and their experience. Once the participants finished creating their stories, we had them complete a post-study survey consisting of SUS and MISCI questions. Finally, a semi-structured interview was conducted to understand the participants' creative process along with their perceptions, preferences, criticisms, and suggestions for ID.8 and the co-creative process. Participants were compensated with a \$40 gift card.

4.3.2 Participants. We recruited 6 participants (3 females, 3 males) via online chat forums. Participants were aged 24 to 25 ($M = 24.5, SD = 0.5$) and had a range of backgrounds and expertise including visual arts, education, public health and electrical engineering. Participants had a range of familiarity with LLMs (neutral=3, unfamiliar=1, very familiar=2),

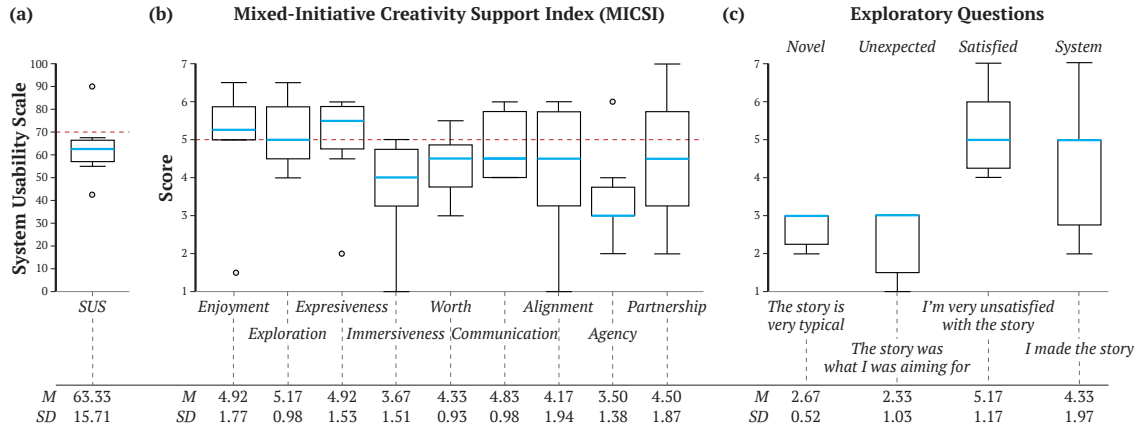


Fig. 7. Results from Study 2: (a) SUS Scores, (b) MICSI Sub-Scale Scores, (c) MISC Exploratory Scores.

familiarity with image/audio generation models (unfamiliar=1, very unfamiliar=5), and familiarity with creating visual content (very familiar=1, familiar=2, unfamiliar=2, very unfamiliar=2).

4.3.3 Results. We provide the quantitative results in Fig. 7 and report on key observations from the analysis of the survey data and transcriptions of the user interviews below.

This evaluation yielded lower SUS scores ($M = 63.33, SD = 15.71$) and Enjoyment scores ($M = 4.92, SD = 1.84$) relative to Study 1 which may have been due to the overhead of launching docker containers through a command line interface to get the system running. Still, participants reported enjoying using the system in the interviews and were satisfied by the stories they were able to author using ID.8 as indicated by the exploratory question regarding story satisfaction (see Fig. 7-c).

P14: “Thank you for letting me use the tool. It was cool and now I have it on my laptop. So I’ll keep using it.”

P15: “I just think that it was really cool and I would actually like to just use it on my own. I thought it was like, I feel like now that we did round one, if we did a round two would be way better, you know, like with what we could create.”

This study resulted in a collection of artistically diverse stories each about 5 minutes in length (see Fig. 6) spotlighting the creative width of ID.8⁶. e.g., P13 created an visually engaging story, see Fig. 8, with captivating audio elements and dramatic narration using ID.8 with a dark, grunge atmosphere with fantastical theme. The creative diversity of the authored stories and the Exploration ($M = 5.17, SD = 0.98$) score, see Fig. 7-b, further establish that ID.8 supports creative exploration reasonably well (DC3).

The Alignment ($M = 4.17, SD = 1.94$) and Communication ($M = 4.83, SD = 0.98$) sub-scale scores indicate a need to enable users to effectively communicate their intent with the co-creative system.

This evaluation supports the findings of Study 1 regarding the need to enhanced sense of collaboration between the user and ID.8; participants again felt that their contributions were equally matched by the system (see Fig. 7-c) but the Partnership ($M = 4.50, SD = 1.87$) and Agency ($M = 3.50, SD = 1.38$) scores spotlight the need for improving

⁶link to google drive folder with sample stories from study 2: <https://tinyurl.com/y3hjswhc>

the collaborative experience towards ID.8 being perceived as a collaborator than just a tool. The low Immersiveness ($M = 3.67, SD = 1.51$) score further underscore the need to facilitate a more immersive co-creative experience.

Users reported spending 7–8 hours exploring the system and producing their stories. It was consistently reported that creating the first few scenes took most of the time; however, as they got used to the system, the creation process became faster. The manifestation of the ID.8’s learning curve and discovery of shortcomings in the co-creative process in Study 2’s less constrained creative setting highlights the value of conducting an open-ended, longer-term, in-the-field evaluation to gain a well-rounded perspective on a system’s usability.

P12: *“By the end, I feel like I was more used to knowing what to do and each scene definitely the time spent became like shorter and shorter.”*

The Knight and The Dragon

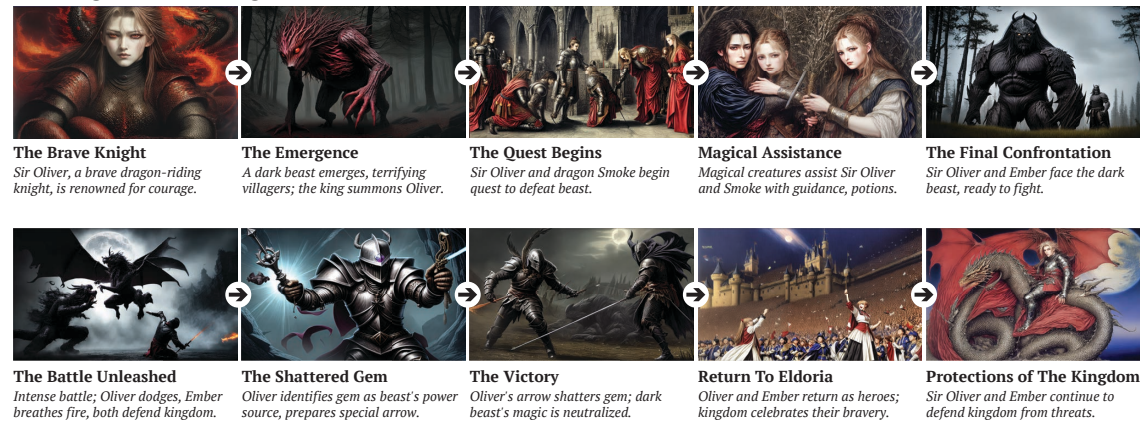


Fig. 8. Scenes depicting the story titled 'The Knight and The Dragon' created by P13 using ID.8.

5 LESSONS LEARNED AND DESIGN GUIDELINES

As generative AI increasingly influences the nature of creative work and content production across domains, it has been suggested that the prominent paradigm of human-AI interaction will be co-creation centered around communicating with generative models, with users supplying higher-level directions to complete complex downstream tasks [33, 39]. There are key design challenges in empowering users to effectively communicate their intent, which will in turn enhance the immersiveness of the co-creation process, strengthen the sense of partnership between the AI and the user, and promote the perception of the AI as a collaborator. Our exploratory findings of participants’ experiences with ID.8 can serve as a proxy for the kinds of multi-agent, multimodal human-AI interactions that are predicted to become the norm in the future of creative work. Here, we provide the key lessons learned from our evaluation of ID.8 regarding the potential and challenges of using generative AI and present a set of synthesized design guidelines (see Table 2) to inform the development of the interaction and workflow architectures of near-future human-AI co-creative systems.

5.1 Users' Mental Models of Generative AI Are Weak

While participants' experiences with ID.8 revealed both strengths and weaknesses of generative AI's capabilities, a significant barrier to effective interaction lies in users' limited mental models of how these systems work, particularly in the area of prompt engineering for content generation.

5.1.1 Addressing Prompting Difficulties with Prompt Templates. Participants appreciated leveraging generative models to create multimedia assets for their stories; specifically, they valued the ability to quickly explore a wide range of creative possibilities and the workflow's efficiency as compared to manual asset creation. However, participants felt limited in their ability to precisely describe the specific assets they wanted to generate. For instance, they encountered difficulties in 1) avoiding generation certain unwanted content (*e.g.*, unnatural anatomy); 2) generating the same character with varying poses, clothing, and facial expressions; and 3) maintaining stylistic consistency across different generated assets. While some participants tried using specific artists' names in their prompts to achieve stylistic consistency, they

Table 2. Design Guidelines for Human-AI Co-Creative Systems

Challenge: Users' Mental Models of Generative AI Are Weak	
Design Guideline 1	Offer predesigned prompt templates with defined fields to streamline the input process and facilitate user intent communication with generative models.
Design Guideline 2	Provide a library of example outputs alongside the prompts that generated them to enhance users' understanding of how to effectively guide generative models.
Design Guideline 3	Leverage LLMs to assist users in crafting prompts that effectively communicate their creative vision to generative models.
Design Guideline 4	Use intuitive and semantically accurate descriptors or proxies for advanced model parameter controls to improve user understanding and increase usage.
Challenge: Generative AI Outputs Biased and Harmful Content	
Design Guideline 5	Integrate safeguard measures such as trigger warnings and automatic content filters against harmful generative outputs to ensure the emotional well-being of users.
Challenge: Strong Human-AI Partnership Requires Intuitive Communication and Dynamic Collaboration	
Design Guideline 6	Accept multiple input modalities to enable intuitive communication of intent.
Design Guideline 7	Support iterative co-creation through dynamic feedback loops.
Design Guideline 8	Provide parallel processing capabilities to facilitate a more dynamic, active co-creation experience.
Challenge: Co-Creative AI Needs a Unified Identity and an Adaptive Workflow	
Design Guideline 9	Establish a centralized AI identity within the system to streamline user experience, enhance the sense of partnership, and increase immersion.
Design Guideline 10	Adapt to individual creative workflows to encourage the exploration of a variety of creative directions.

reported instances where the generated image would disproportionately reflect the likeness or subject matter focus of a given artist. The struggle to effectively guide the generative models to produce desired outputs is also reflected in the suboptimal MISCI subscale scores for Communication from both studies and the low Expressiveness and Alignment MISCI sub-scale scores from Study 2 (see Fig. 5 and Fig. 7).

P11: *“Visuals generation, like it was gorgeous in a lot of the cases, but it wouldn’t get what I wanted or like it would give me weird things like two-headed animals.”*

Our observations indicate that the process of prompting generative models is not intuitive enough to the average person; there is a stark need to simplify the art of prompt engineering [64]. Providing prompt templates (e.g., *[Medium][Subject][Artist(s)][Details][Image repository support]* [38]) with a wide set of prompt modifiers (e.g., style specifiers, quality boosters [38, 41]) may be an effective approach toward building stronger user mental models of how to communicate intent and creative vision to generative AI models. While research exists on prompting techniques for text-to-image and LLMs, there is a significant gap in how to prompt generative models for other modalities (e.g., audio, video, speech); therefore, future work should focus on creating modality-specific prompt templates to fully leverage the co-creative potential of generative AI in multimodal settings.

Design Guideline: *Offer predesigned prompt templates with defined fields (e.g., *[Medium][Subject][Artist(s)][Details]*) to streamline the input process and facilitate user intent communication with generative models.*

5.1.2 Building Intuition on Model Capabilities Using a Prompt-Output Pair Library. Generative models have very impressive capabilities, but their outputs are greatly dependent on the prompts that they are provided with. We noticed that the difficulties participants experienced while prompting the models resulted in misunderstanding what the models are capable of generating. For instance, participants reported having trouble generating specific animals (e.g., narwals, dragons) and maintaining stylistic consistency across generations; similarly, even though participants were generally successful in prompting Leela (ChatGPT), some desired more emphasis on character development, the inclusion of thematic elements such as a “moral of the story” and the use of simpler language for younger target audiences. Interestingly, ChatGPT is capable of making these changes when asked further, spotlighting a lack of coherent understanding of the model’s capabilities.

P12: *“For me, it was really hard making consistent images. And I felt like it was really hard creating like a cohesive image group. like the cave scene literally looked like somebody took a picture of a cave and the next scene has some anime girlies fighting a thing. It just feels super shocking to someone who like works with visual stuff. When you want everything to feel like it’s a one story, you know, not like different parts of some weird thing.”*

P13: *“You don’t see a lot of dragon in my story because [the visuals generator] just couldn’t do dragon. Its closest estimation was like a horse.”*

In contrast, expert users of generative AI are able to guide generative models to produce outputs that are consistent in style⁷ and spatial location⁸—and they can even produce dragons⁹. Our observations highlight the gaps in participants’ understanding of what “black-box” generative models like ChatGPT are capable of achieving, highlighting a need to build better user intuition as to these models’ capabilities and how to guide the models to materialize a creative vision [64].

⁷<https://twitter.com/chaseleantj/status/1700499968690426006>

⁸<https://twitter.com/Salmaaboukarr/status/1701215610963665067>

⁹https://twitter.com/art_hiss/status/1701623410848096551

Providing a diverse library of generated outputs coupled with the prompts used to generate them, as suggested by some of our study participants, may allow for enhancing users’ mental models of how to guide these AI models to generate the desired content [60]. This may be particularly effective in the case of models that are able to accept inputs beyond just text.

Design Guideline: *Provide a library of example outputs alongside the prompts that generated them to enhance users’ understanding of how to effectively guide generative models.*

5.1.3 Leveraging LLMs to Enable Effective Prompt Engineering. Recent studies have demonstrated LLMs’ abilities in engineering effective prompts for generative models [10, 16, 18, 23, 63, 71]. We observed that while attempting to engineer effective prompts to generate artifacts that matched their creative vision, some participants turned to Leela (ChatGPT) for assistance in crafting more detailed prompts. They found that Leela was useful in helping them get closer to their intended output by providing helpful terms that could guide the diffusion models to produce more desired outputs, although the participants also mentioned that its prompts were overly verbose.

P4: *“I guess also being familiar with the types of styles or types of animation and drawing, like the names of certain styles. Like I’m not super familiar with the names of certain styles. So I wouldn’t know how to describe that besides, like, I don’t know, make it look like my fairy godparents.”*

P11: *“It took a while for the generation, I guess, to get what I really wanted until I used Leela. She described, used more descriptive words, and then copying it from there actually helped a lot. So, like, I didn’t use that feature at the beginning as much . . . copying from there, even though the prompts were really long from Leela, picking up some descriptions was really helpful.”*

Our observations suggest there is an opportunity for further research into how best to design effective collaborations between LLMs and users for the purpose of jointly crafting prompts. Our prompt for Leela—instructing it to help provide the user with descriptive prompts—was relatively simple and yet yielded helpful outcomes; future work should focus on how to facilitate constructive conversations that ensure the LLM gains a clearer understanding of the user’s creative vision [2, 59]. Moreover, exploring this collaborative interaction could be a fruitful avenue for enhancing creative exploration and imagination, as LLMs may promote new stylistic or functional directions to explore via their suggested prompts.

Design Guideline: *Leverage LLMs to assist users in crafting prompts that effectively communicate their creative vision to generative models.*

5.1.4 Simplifying Model Parameter Terminology. Although generative models are able to generate an increasingly impressive and practically infinite range of content and their capabilities may continue to span many more domains, a key limitation is the lack of easily understandable controls for the novice user [33]. Consequently, we observed that none of our study participants utilized the advanced model options, such as self-attention mechanisms, denoising steps, or random seeds; using these advanced controls could help achieve better control over the models’ outputs. The absence of engagement with these advanced controls was attributed to a lack of familiarity with these technical terms, suggesting a need for more user-friendly terminology to effectively explain the use and impact of these controllable parameters. To further elaborate, the term “denoising steps” could be simplified to “boost clarity” along with a tooltip or brief description that clarifies its purpose: “The ‘boost clarity’ option helps eliminate random noise from the model’s output, making it cleaner and more focused, but requiring longer generation time.” The key hurdle lies in simplifying the language without sacrificing an accurate representation of how these parameters can impact the model’s behavior.

Providing semantically equivalent proxies for advanced model parameters in form of sliders or gestural inputs may also be a viable avenue to democratize advanced model control parameters [13, 34].

Design Guideline: *Use intuitive and semantically accurate descriptors or proxies for advanced controls to improve user understanding and increase usage.*

5.2 Generative AI Outputs Biased and Harmful Content

While the potential for the positive impact of generative-AI-enabled co-creative systems should not be ignored, there are critical safety and bias considerations that must be addressed [17]. We learned that it is worryingly easy to get generative models to output biased and/or inappropriate content. The bias built into the training data of these models should be addressed and safeguards must be installed. Our study participants encountered harmful content in the form of biased and unsafe outputs generated by the image generation model; participants reported pornographic, gory, and racially biased outputs generated by the models.

Underscoring the concern about inappropriate and harmful outputs, one participant reported a specific experience that was problematic in terms of both sexualization and racial stereotyping:

P12: *“The main frustration I had was working with a story about ninja princesses, which generated either generic Caucasian princesses or like super fetishized Asian women who had no clothing on.”*

Bias was not confined to sexual or violent content. A participant reported an instance of racial bias:

P9: *“I asked it to generate, like, a city after an earthquake and all the cities that are generated were out of almost like a third world country. Then, and then, when I asked to put people in it, it was really like all the people were people of color.”*

The ethical considerations in this context extend beyond mere content moderation or output filtering; they also concern user safety and emotional well-being. Exposure to such biased or unsafe content can be especially distressing when the user belongs to the stereotyped or marginalized group in question. One participant elaborated on this emotional toll:

P12: *“I was frustrated at the AI, I was mostly frustrated at like knowing that what makes the AI work is like the bank of information on the internet and, like, what is mostly available on the internet. And because it was producing such, like, fetishized, stereotyped images, I knew that it was because there’s such a large amount of that, like in the world, on the internet. So that was making me most frustrated. Because it kind of felt like I was being confronted with it in like a weird way.”*

This reiterates the need for maintaining the human-in-control, AI-in-the-loop design philosophy of our system; however, it also highlights a shortcoming of our current design and emphasizes the need to safeguard human creators without hindering their creative autonomy. The emotional stress cited by participants, particularly when their own identity was implicated, underscores the urgency of resolving these ethical issues. This experience elucidates not only the importance of implementing robust filtering mechanisms but also points to a crucial need for more ethical considerations in the design and training of generative models. The impact of these unsafe and biased outputs on users’ emotional states also highlights the importance of integrating emotional safety measures, perhaps through the use of trigger warnings or other alert systems, as part of a holistic approach to system design.

Design Guideline: *Integrate safeguard measures such as trigger warnings and automatic content filters against harmful generative outputs to ensure the emotional well-being of users.*

5.3 Strong Human-AI Partnership Requires Intuitive Communication and Active Collaboration

Our evaluation of ID.8 revealed nuanced aspects of human-AI interaction in co-creative settings, bringing to light four critical areas requiring attention for more effective co-creation: intuitive communication of user intent, iterative collaboration for alignment of objectives, parallel task execution for active engagement, and the necessity of an immersive co-creative system experience.

5.3.1 Supporting Multimodal Input for Intuitive Communication During Co-Creation. Our study participants indicated that text-based communication can be a limitation in effective human-AI collaboration; participants expressed a desire to interact with generative models through multiple input modalities (e.g., sample images, sketches) to better express their creative vision. The suboptimal Expressiveness and Alignment MISCI subscale scores from Study 2 underscore the limitations of a text-only communication pathway in a multimodal co-creative process.

P5: *“If you find a similar kind of image, there should be an option for us to tell the AI... yeah, this is kind of something similar, I want more image like this coming up.”*

Recent human-AI co-creative systems (e.g., Talebrush [13], BrainFax [57]) have demonstrated how more intuitive controls (e.g., sketch-based inputs) can facilitate more natural human-AI communication, empowering users to materialize their creative vision more easily [42, 65]. Co-creative systems should either provide users access to a variety of models (e.g., Uni-ControlNet [70], SketchyGAN [11]) with a range of input modalities or integrate methods to convert the user’s multimodal input to a form accepted by the generative model, thereby enabling more intuitive communication of the user’s creative intent.

Design Guideline: *Accept multiple input modalities to enable intuitive communication of intent.*

5.3.2 Support Iterative Collaboration to Improve Human-AI Alignment. A noticeable difference between participants’ perceptions of Leela (ChatGPT) relative to that of the asset generation models was due to the lack of ability to iterate on the generated assets.

P15: *“So I typed in to Leela that I wanted a story about a penguin who learns how to share and then she spit out a story to me that I liked, but I wanted it to be a little bit more complex. And I didn’t like the name she chose. So I asked for a different name and then I asked for her to add a couple different elements into it and then she spat out a story that I really liked.”*

P12: *“And I feel like once images were created, if I kept adding to the prompt and regenerating, it felt as if it was already on one path. So in order to do something new, I had to like, change the prompt like entirely and shift things up. And because of that, I feel like I had a hard time creating an image that I was really satisfied with. I think a lot of the times it was like, OK, I think this is gonna be the best that is gonna come out. So I was like, just whatever, that’s fine.”*

These observations suggest that users may prefer co-creative models that are able to iterate on their outputs based on user feedback until they produce an acceptable output—as opposed to models that only produce sequentially independent outputs. The lack of such iterative co-creation may also lead to a less immersive collaboration experience; the suboptimal Immersiveness and Partnership MICSIS subscale scores from both our studies further indicate a need to improve the co-creative process through an enhanced sense of collaboration. Incorporating the ability to iterate on outputs sequentially may require further research on architectural changes to generative models, but could significantly improve user satisfaction and creative outcomes. Co-creative systems should be designed to allow for iterative output

based on real-time user feedback; this could involve mechanisms that let users adjust the parameters of generated artifacts without starting from scratch or review and amend intermediate outputs before the final artifact is generated.

Design Guideline: *Support iterative co-creation through dynamic feedback loops.*

5.3.3 *Parallel Processing Leads to a Sense of Active Collaboration.* State-of-the-art generative models, while powerful, are resource-intensive and often require lengthy generation times on commercial hardware; this poses a significant challenge for designing immersive, effective human-AI co-creative systems. We observed that participants were particularly frustrated with the long wait times for asset generation; they expressed a desire for features that would allow them to queue or minimize the generation process, enabling them to work on other aspects of their projects in parallel. Participants emphasized that these wait times significantly affected their co-creative experience, making it feel less like an active collaboration and more like a turn-taking exercise. This is reflected in the low Partnership MICSI subscale scores from Study 2.

P8: *“I wasn’t sure what to do when I was waiting for the content to be generated. It would be nice if there would be a way to add a generation to a queue. And then you could still be doing something on the screen while you’re waiting for it to be generated. I think that would help with feeling like it was more of an active collaboration. I mean, right now it did feel like there was a feeling of collaboration to it. But it was sort of like we trade off who’s the one working on it rather than maybe working on something more actively together.”*

Design Guideline: *Provide parallel processing capabilities to facilitate a more dynamic, active co-creation experience.*

5.4 Co-Creative AI Needs a Unified Identity and an Adaptive Workflow

In our evaluation of ID.8, we observed that the system’s fragmented set of generative models and rigid authoring workflow weakened the perception of collaboration, highlighting the need for a centralized AI presence capable of dynamically adapting to individual creative workflows.

5.4.1 *A Central AI Identity Helps Foster a Stronger Sense of Collaboration.* The nature in which the AI manifests itself in the co-creative system has critical implications on whether the system is perceived as a tool or a collaborator; the discrete chunks AI support exists in the co-creative process in ID.8—represented by different models or modules for various tasks—disrupts the flow of the creative process. We found that participants desired more integrated AI support throughout the creation process; for instance, one participant noted a desire to brainstorm more closely with the AI to the extent of even generating appropriate animations for a character. Participants also noted how Leela (ChatGPT) did not seem integrated into the rest of the workflow, further indicating a desire for a more cohesive interaction with a “single agent” instead of an amalgamation of models. The suboptimal MICSI scores for the Partnership and Agency subscales reflect the participants’ perceptions of ID.8 as more of a tool than a collaborator.

P12: *“I forgot that Leela was an option ... Because to me, it felt like trial and error, you know what I mean? So it felt like one more step of interacting with a separate thing to get prompts, to interact ...”*

Creating a central identity for the AI in co-creative system, either through a wrapper that coordinates a cohesive interaction experience with a range of generative models or through a foundational model capable of executing all the tasks required in the co-creative domain, would lead to better sense of partnership and immersion in the co-creative process.

Design Guideline: *Establish a centralized AI identity within the system to streamline user experience, enhance the sense of partnership, and increase immersion.*

5.4.2 *Encourage Creative Exploration While Adapting to Creative Process.* Generative AI in co-creative systems cannot only help materialize users’ creative visions, but may also help expand their creative imagination [37]. Our study participants reported that the AI changed their minds about what they wanted to do with their stories several times and in various manners (e.g., by suggesting alternative plot lines, by frustrating them with an inability to generate the desired type of output, by showing new ways of envisioning characters, etc). The positive MICSI scores for the Exploration subscale from both our studies further highlight the utility of generative models in encouraging creative exploration in one way or another.

P7: “[While creating the storyline] at the very first I thought, oh, this magic boy might be a naughty boy. He might play some tricks, you know, do some bad things. But when I type it, the chatbot told me perhaps he could do some good things. So, yeah, [it] totally changed my mind”

P15: “I think like, once you see the script and then you get inspiration of what you envision and you want it to match exactly that. You have to kind of like play around a bit. But I found that a few times I did find like, oh, I was like, okay, this is perfect. This is exactly exactly what I was thinking. And then other times when I couldn’t quite get what I wanted, it kind of just took me in a new direction. I was like, oh, okay, I didn’t envision it like this, but this is really cool.”

P14: “[While] creating images, sometimes whatever you describe is not what it spits out. But at the same time, it could end up leading you to create new ideas or, like, give you more inspiration that oh, Okay, the character could look like this and then you try to describe it in a better way.”

P16, who did not think the system helped boost creativity, suggested that perhaps if the system suggested a variety of styles to choose from during a “sandbox”-based asset creation process prior to scene authoring, it would help users be more creative. This indicates that to be an effective co-creative partner and optimally encourage creative exploration, co-creative agents must dynamically adapt to individual differences in the creative process; for example, they should support a non-linear workflow of ideation and creation, rather than simply following a linear pipeline (e.g., from storyline crafting to asset creation to visual story construction).

Design Guideline: *Support individual creative workflows to encourage the exploration of a variety of creative directions.*

6 LIMITATIONS AND FUTURE WORK

Our evaluations yielded insightful findings for human-AI co-creative systems in a multi-modal creative process. However, the study remains limited due to being conducted with a controlled task disconnected from an organic, creative workflow in a real workplace with more realistic incentives. Studying human-AI co-creative systems in a real-world setting may elucidate further opportunities and challenges and help elaborate design guidelines for co-creative systems. Moreover, the results of our study could benefit from being extended through evaluating the experience of experts in the creative domain in using a co-creative system such as ID.8; such an extension would help clarify how AI should adapt to different individual creative processes and how the creative support that it provides would need to be adjusted.

AI supporting human teams in creative domains is yet another paradigm that may emerge as generative AI becomes more and more prevalent. There is currently a lack of robust understanding of what roles AI can play in this partnership and what interaction strategies may be appropriate in a multi-human-AI team conducting a multimedia co-creative task. Generating interactive, visual content is typically a team-effort, especially when producing content for sensitive uses cases such as psycho-educational content; studying team-based human-AI collaboration may help establish more robust design guidelines for co-creative system in this real-world context.

Lastly, we designed and engineered ID.8 to support the end-to-end authoring of visual stories supported by generative AI. Visual stories have a wide range of use cases such as educational content and psychotherapeutic aids. Evaluating how well the outputs of ID.8 support such use cases can help further establish its value as a useful tool along with serving as a platform to study human-AI co-creativity.

7 CONCLUSION

Generative AI has the potential to lower barriers to creative expression and visual story generation. Our work contributes ID.8, an open-source, end-to-end visual story authoring system that integrates the state-of-the-art generative models in the creative workflow to lower barriers. Our evaluation demonstrates the potential of human-AI co-creative systems such as ID.8 and elucidates areas for improvement along with challenges users face when collaborating with generative AI in creative work. Our findings inform design guidelines for the future human-AI co-creative systems.

REFERENCES

- [1] Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara J Martin, and Mark O Riedl. 2020. Story realization: Expanding plot events into sentences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 7375–7382.
- [2] Seunggho Baek, Hyerin Im, Jiseung Ryu, Juhyeong Park, and Takyoon Lee. 2023. PromptCrafter: Crafting Text-to-Image Prompt through Mixed-Initiative Dialogue with LLM. *arXiv preprint arXiv:2307.08985* (2023).
- [3] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies* 4, 3 (2009), 114–123.
- [4] Weizhen Bian, Yijin Song, Nianzhen Gu, Tin Yan Chan, Tsz To Lo, Tsun Sun Li, King Chak Wong, Wei Xue, and Roberto Alonso Trillo. 2023. MoMusic: A Motion-Driven Human-AI Collaborative Music Composition and Performing System. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 16057–16062.
- [5] Andrew S Bradlyn, Ivan L Beale, and Pamela M Kato. 2003. Psychoeducational interventions with pediatric cancer patients: Part I. Patient information and knowledge. *Journal of child and family studies* 12 (2003), 257–277.
- [6] John Brooke. 1996. Sus: a “quick and dirty” usability. *Usability evaluation in industry* 189, 3 (1996), 189–194.
- [7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [8] Daniel Buschek, Lukas Mecke, Florian Lehmann, and Hai Dang. 2021. Nine potential pitfalls when designing human-ai co-creative systems. *arXiv preprint arXiv:2104.00358* (2021).
- [9] Tony C Caputo. 2003. *Visual Storytelling: the art and technique*. Watson-Guption Publications.
- [10] Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. *arXiv preprint arXiv:2305.14724* (2023).
- [11] Wengling Chen and James Hays. 2018. Sketchygan: Towards diverse and realistic sketch to image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9416–9425.
- [12] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).
- [13] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: visual sketching of story generation with pretrained language models. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–4.
- [14] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and Controllable Music Generation. *arXiv preprint arXiv:2306.05284* (2023).
- [15] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341* (2020).
- [16] Wala Elsharif, James She, Preslav Nakov, and Simon Wong. 2023. Enhancing Arabic Content Generation with Prompt Augmentation Using Integrated GPT and Text-to-Image Models. In *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*. 276–288.
- [17] Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, et al. 2022. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1747–1764.
- [18] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. 2023. Text-to-Audio Generation using Instruction-Tuned LLM and Latent Diffusion Model. *arXiv preprint arXiv:2304.13731* (2023).
- [19] Roberto Gozalo-Brizuela and Eduardo C Garrido-Merchán. 2023. A survey of Generative AI Applications. *arXiv preprint arXiv:2306.02781* (2023).

- [20] Ariel Han and Zhenyao Cai. 2023. Design implications of generative AI systems for visual storytelling for young learners. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. 470–474.
- [21] Daphne Ippolito, Ann Yuan, Andy Coenen, and Sehmon Burnam. 2022. Creative writing with an ai-powered writing assistant: Perspectives from professional writers. *arXiv preprint arXiv:2211.05030* (2022).
- [22] Amir Jahanlou and Parmit K Chilana. 2022. Katika: An End-to-End System for Authoring Amateur Explainer Motion Graphics Videos. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [23] Hyeonho Jeong, Gihyun Kwon, and Jong Chul Ye. 2023. Zero-shot generation of coherent storybook from plain text story using diffusion models. *arXiv preprint arXiv:2302.03900* (2023).
- [24] Nam Wook Kim, Nathalie Henry Riche, Benjamin Bach, Guanpeng Xu, Matthew Brehmer, Ken Hinckley, Michel Pahud, Haijun Xia, Michael J McGuffin, and Hanspeter Pfister. 2019. Datatoon: Drawing dynamic network comics with pen+ touch interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643* (2023).
- [26] Tim Knapp. 2023. Situating Large Language Models Within the Landscape of Digital Storytelling. In *Proceedings of the MEI: CogSci Conference*, Vol. 17.
- [27] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2022. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352* (2022).
- [28] Tomas Lawton, Kazjon Grace, and Francisco J Ibarrola. 2023. When is a Tool a Tool? User Perceptions of System Agency in Human–AI Co-Creative Drawing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 1978–1996.
- [29] Tomas Lawton, Francisco J Ibarrola, Dan Ventura, and Kazjon Grace. 2023. Drawing with Reframer: Emergence and Control in Co-Creative AI. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 264–277.
- [30] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).
- [31] Cun Li, Jun Hu, Bart Hengeveld, and Caroline Hummels. 2019. Story-me: design of a system to support intergenerational storytelling and preservation for older adults. In *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion*. 245–250.
- [32] Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang, Fu Li, Haocheng Feng, Errui Ding, and Jingdong Wang. 2023. VideoGen: A Reference-Guided Latent Diffusion Approach for High Definition Text-to-Video Generation. *arXiv preprint arXiv:2309.00398* (2023).
- [33] Vivian Liu and Lydia B Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–23.
- [34] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J Cai. 2020. Novice-AI music co-creation via AI-steering tools for deep generative models. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.
- [35] Kristijan Mirkovski, James E Gaskin, David M Hull, and Paul Benjamin Lowry. 2019. Visual storytelling for improving the comprehension and utility in disseminating information systems research: Evidence from a quasi-experiment. *Information Systems Journal* 29, 6 (2019), 1153–1177.
- [36] Eric Mörth, Stefan Bruckner, and Noeska N Smit. 2022. ScrollyVis: Interactive visual authoring of guided dynamic narratives for scientific scrollytelling. *IEEE Transactions on Visualization and Computer Graphics* (2022).
- [37] Jonas Oppenlaender. 2022. The creativity of text-to-image generation. In *Proceedings of the 25th International Academic Mindtrek Conference*. 192–202.
- [38] Jonas Oppenlaender. 2022. A taxonomy of prompt modifiers for text-to-image generation. *arXiv preprint arXiv:2204.13988 2* (2022).
- [39] Jonas Oppenlaender, Rhema Linder, and Johanna Silvennoinen. 2023. Prompting ai art: An investigation into the creative skill of prompt engineering. *arXiv preprint arXiv:2303.13534* (2023).
- [40] Hiroyuki Osone, Jun-Li Lu, and Yoichi Ochiai. 2021. BunCho: ai supported story co-creation via unsupervised multitask learning to increase writers’ creativity in japanese. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–10.
- [41] Nikita Pavlichenko and Dmitry Ustalov. 2023. Best prompts for text-to-image models and how to find them. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2067–2071.
- [42] Han Qiao, Vivian Liu, and Lydia Chilton. 2022. Initial images: using image prompts to improve subject representation in multimodal ai generated art. In *Proceedings of the 14th Conference on Creativity and Cognition*. 15–28.
- [43] Chia Yi Quah and Kher Hui Ng. 2022. A systematic literature review on digital storytelling authoring tool in education: January 2010 to January 2020. *International Journal of Human–Computer Interaction* 38, 9 (2022), 851–867.
- [44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125 1, 2* (2022), 3.
- [45] Jeba Rezwana and Mary Lou Maher. 2022. Identifying ethical issues in ai partners in human-ai co-creation. *arXiv preprint arXiv:2204.07644* (2022).
- [46] Jeba Rezwana, Mary Lou Maher, and Nicholas Davis. 2021. Creative PenPal: A Virtual Embodied Conversational AI Agent to Improve User Engagement and Collaborative Experience in Human-AI Co-Creative Design Ideation.. In *IUI Workshops*.
- [47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

- [48] Carolina Beniamina Rutta, Gianluca Schiavo, and Massimo Zancanaro. 2019. Comic-based digital storytelling for self-expression: an exploratory case-study with migrants. In *Proceedings of the 9th International Conference on Communities & Technologies-Transforming Communities*. 9–13.
- [49] Nisha Simon and Christian Muise. 2022. TattleTale: Storytelling with Planning and Large Language Models. In *ICAPS Workshop on Scheduling and Planning Applications*.
- [50] Helena Romano Snyder and Israel Colon. 1988. Foreign language acquisition and audio-visual aids. *Foreign Language Annals* 21, 4 (1988), 343–348.
- [51] Michelle Scalise Sugiyama. 2001. Narrative theory and function: Why evolution matters. *Philosophy and literature* 25, 2 (2001), 233–250.
- [52] Sangho Suh, Sydney Lamorea, Edith Law, and Leah Zhang-Kennedy. 2022. PrivacyToon: Concept-driven Storytelling with Creativity Support for Privacy Concepts. In *Designing Interactive Systems Conference*. 41–57.
- [53] Lingyun Sun, Pei Chen, Wei Xiang, Peng Chen, Wei-yue Gao, and Ke-jun Zhang. 2019. SmartPaint: a co-creative drawing system based on generative adversarial networks. *Frontiers of Information Technology & Electronic Engineering* 20, 12 (2019), 1644–1656.
- [54] Ben Swanson, Kory Mathewson, Ben Pietrzak, Sherol Chen, and Monica Dinalescu. 2021. Story centaur: Large language model few shot learning as a creative writing tool. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. 244–256.
- [55] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
- [56] Kelly LA van Bindsbergen, Hinke van der Hoek, Marloes van Gorp, Mike EU Ligthart, Koen V Hindriks, Mark A Neerinx, Tanja Alderliesten, Peter AN Bosman, Johannes HM Merks, Martha A Grootenhuis, et al. 2022. Interactive education on sleep hygiene with a social robot at a pediatric oncology outpatient clinic: feasibility, experiences, and preliminary effectiveness. *Cancers* 14, 15 (2022), 3792.
- [57] Mathias Peter Verheijden and Mathias Funk. 2023. Collaborative Diffusion: Boosting Designerly Co-Creation with Generative AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–8.
- [58] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. 2023. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.
- [59] Yunlong Wang, Shuyuan Shen, and Brian Y Lim. 2023. RePrompt: Automatic Prompt Editing to Refine AI-Generative Art Towards Precise Expressions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–29.
- [60] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2022. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896* (2022).
- [61] Justin D Weisz, Michael Muller, Jessica He, and Stephanie Houde. 2023. Toward general design principles for generative AI applications. *arXiv preprint arXiv:2301.05578* (2023).
- [62] Qiang Wu, Baixue Zhu, Binbin Yong, Yongqiang Wei, Xuetao Jiang, Rui Zhou, and Qingguo Zhou. 2021. ClothGAN: generation of fashionable Dunhuang clothes using generative adversarial networks. *Connection Science* 33, 2 (2021), 341–358.
- [63] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023. Large Language Models as Optimizers. *arXiv preprint arXiv:2309.03409* (2023).
- [64] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.
- [65] Chengzhi Zhang, Weijie Wang, Paul Pangaro, Nikolas Martelaro, and Daragh Byrne. 2023. Generative Image AI Using Design Sketches as input: Opportunities and Challenges. In *Proceedings of the 15th Conference on Creativity and Cognition*. 254–261.
- [66] Chao Zhang, Cheng Yao, Jiayi Wu, Weijia Lin, Lijuan Liu, Ge Yan, and Fangtian Ying. 2022. StoryDrawer: A Child-AI Collaborative Drawing System to Support Children's Creative Visual Storytelling. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [67] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. 2023. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909* (2023).
- [68] Lvmin Zhang and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543* (2023).
- [69] Yixiao Zhang, Gus Xia, Mark Levy, and Simon Dixon. 2021. COSMIC: A Conversational Interface for Human-AI Music Co-Creation. In *NIME 2021*. PubPub.
- [70] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. 2023. Uni-ControlNet: All-in-One Control to Text-to-Image Diffusion Models. *arXiv preprint arXiv:2305.16322* (2023).
- [71] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910* (2022).