WorldSmith: Iterative and Expressive Prompting for World Building with a Generative AI

Hai Dang
hai.dang@uni-bayreuth.de
University of Bayreuth & Autodesk Research
Bayreuth, Bavaria, Germany

Frederik Brudy
frederik.brudy@autodesk.com
Autodesk Research
Toronto, Ontario, Canada

George Fitzmaurice
george.fitzmaurice@autodesk.com
Autodesk Research
Toronto, Ontario, Canada

Fraser Anderson
fraser.anderson@autodesk.com
Autodesk Research
Toronto, Ontario, Canada

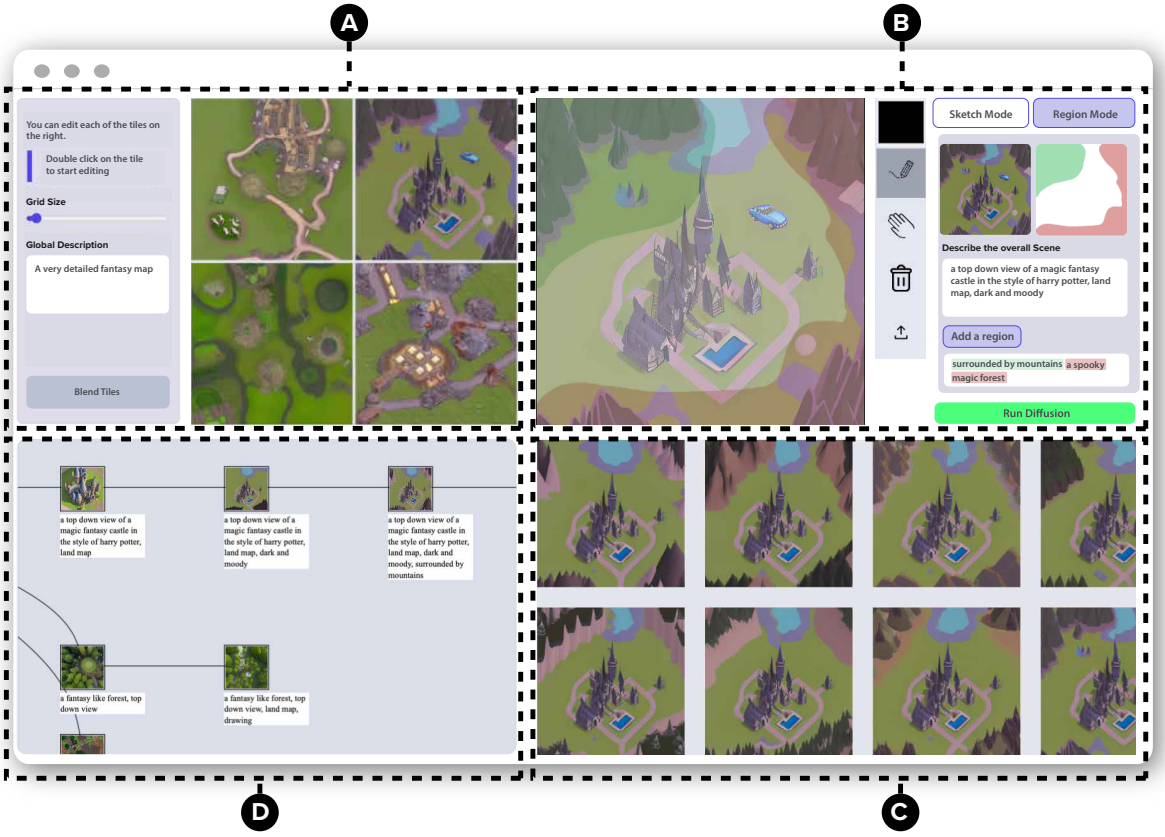

Figure 1: The workflow and high-level interface of WorldSmith. The user selects one of the four image tiles in the *Global Tile View* (A), and iteratively edits this tile with text prompts, sketching, and region-painting tools available through the *Detailed Tile Editor* (B). All generated images are collected in the *Results View* (C) which allows the users to re-use previous image assets. Furthermore, the *Tree View* automatically captures each new image generation request (D). After creating all tiles the user returns to the *Global Tile View* to blend all tiles into a single image.

UIST '23, October 29–November 01, 2023, San Francisco, CA, USA

ABSTRACT

Crafting a rich and unique environment is crucial for fictional world-building, but can be difficult to achieve since illustrating a world from scratch requires time and significant skill. We investigate the

ACM ISBN 979-8-4007-0132-0/23/10…$15.00
https://doi.org/10.1145/3586183.3606772

use of recent multi-modal image generation systems to enable users iteratively visualize and modify elements of their fictional world using a combination of text input, sketching, and region-based filling. WorldSmith enables novice world builders to quickly visualize a fictional world with layered edits and hierarchical compositions. Through a formative study (4 participants) and first-use study (13 participants) we demonstrate that WorldSmith offers more expressive interactions with prompt-based models. With this work, we explore how creatives can be empowered to leverage prompt-based generative AI as a tool in their creative process, beyond current "click-once" prompting UI paradigms.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

Multi-modal image generation, Fictional world-building, AI-assisted creativity

ACM Reference Format:
Hai Dang, Frederik Brudy, George Fitzmaurice, and Fraser Anderson. 2023. WorldSmith: Iterative and Expressive Prompting for World Building with a Generative AI. In *The 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), October 29–November 01, 2023, San Francisco, CA, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3586183.3606772

1 INTRODUCTION

Fictional world-building is the process of constructing a fictional universe with its unique history, geography, culture, and rules [19]. In his seminal essay - *On Fairy Stories* - Tolkien emphasizes the crucial role of an *imagined place* for an engaging fantasy plot. This observation extends into the present, where interesting and intricate worlds are essential for the success of games, movies, and other forms of entertainment.

There are many reasons why people consume fiction with imaginary worlds. While some just enjoy the creative challenge of conceiving an interesting world itself, others appreciate the enormous vibrant community [62]. Forms of engagement include writing fanfiction [65], developing indie games [18], or running the popular table-top game Dungeon and Dragons (DnD)[29].

Fictional worlds can be conveyed through various media, and creating visual artwork that depicts part of the imagined world is one of them. Visually materializing the world helps others find common ground when discussing and collaborating [40]. However, illustrating these worlds is time-consuming and particularly difficult for novice world builders, who often lack the artistic skill or the experience to create coherent worlds. But even experienced world builders must invest significant time and effort into materializing their ideas.

With current illustration tools such as Adobe Photoshop or Adobe Illustrator, users often have to make many fine-grained edits to create their world. This is time-consuming and requires a high level of expertise. To relieve users from designing on the micro-level, many domain-specific tools have been developed, such as procedural terrain generation tools [3, 52, 58], which algorithmically generate varied terrains based on a fixed set of rules or grammar [59, 71]. Other tools have been developed to support the creation of 3D worlds [73, 82] and game levels [72]. However, these tools still operate procedurally, preventing users from defining their world using a high-level semantic description.

Recent image generation models such as Dall-E [60], Stable Diffusion [63], Imagen [64], and Midjourney [53] are now capable of generating high-quality images based on simple natural language text prompts to guide the image generation process semantically. Thus prompting has evolved as a new interaction paradigm with the advent of large pre-trained text [14] and image generation models.

Prompt-based image generation models have become increasingly popular, but the prevailing interaction paradigm with these models is limited to a "click-once" text prompt interface. This approach assumes that users can provide a complete and accurate description of the desired visual imagery upfront. However, world-building is an iterative process, and this simplistic approach may not be adequate [79]. More expressive techniques are needed to interact with these models to address this challenge. One promising approach is incorporating additional input modalities, such as sketching or other graphical interfaces, to allow users to convey their design. However, the impact of such multi-modal systems on the world-building process and the behavior of users when defining prompts for generative AI models remains an open question.

To this end, we designed and built WorldSmith (Figure 1), a tool to support world-builders generate an image of their envisioned fictional worlds through multi-modal inputs, including text input, sketching, and region painting. To accommodate their iterative and piece-by-piece workflow, WorldSmith was designed to reinforce two key concepts 1) hierarchical generation of multiple image tiles and 2) layered editing of individual image tiles. To evaluate the utility and to observe users' prompting behavior with WorldSmith, we conducted a first-use study with 13 participants.

In summary, we contribute the following:

- WorldSmith, a multi-modal tool that enables users to iteratively design and refine complex fictional worlds using layered editing and hierarchical compositions with a prompt-based model that uses text, sketches, and region masks as inputs.
- Insights from a formative and first-use study, demonstrating how WorldSmith facilitates interactive prompting with text input and additionally with non-textual interaction such as sketching and region painting, to disambiguate text prompts for generative AI.

2 RELATED WORK

This work draws upon the domains of world-building, scene generation and human-AI co-creativity.

2.1 Image Synthesis Techniques

To facilitate working with multiple generated images, WorldSmith uses various image composition techniques, including inpainting, which involves filling in missing or damaged areas of an image by generating plausible content based on the surrounding context [76]. Inpainting has been applied to automatically colorize rough

sketches [66] and remove objects from photographs [12]. Another related technique is outpainting, which generates new content beyond the boundaries of an image [75]. Pre-trained image models have also been investigated for their ability to perform visual conceptual blending [23], which involves blending visual concepts to generate new content, such as an "amphibious vehicle" resulting from blending "a boat" and "a bus." Other research has explored blending for creating symbols [9] and for sketches [13, 37]. Multiple inputs, including fully segmented images with corresponding annotations [4, 22], can be used to synthesize images. Motivated by the world-building workflow, we investigate how various image synthesis techniques work together to support users in their world-building process. Specifically, we blend multiple image tiles to create a larger composition, generating new content beyond the boundaries of individual images while staying within the boundaries of the target set as a whole.

2.2 Prompting Pretrained Generative Models

Recent developments in natural language processing have shown that Pre-trained Large Language Models (LLMs) can solve multiple tasks without the need for specific training for each task. This can be achieved by using text prompts in natural language, as demonstrated by Brown et al. [7]. Generating effective text prompts is a challenging task, not just for generating text [45], but also for generating images. Although the internet community has developed several prompting strategies to create more targeted images, such as incorporating resolution-related terms like *4k* or *Unreal Engine*, recent research has proposed techniques to automatically refine these prompts through prompt engineering [30, 35, 44, 78] or interactive methods [46, 47]. While, many interactive prompt-based tools only support uni-modal input [11, 47, 67], two recent surveys [39, 40] independently called for also exploring multi-modal affordances of prompt-based models. To this end, Liu et al. [48] designed 3DALL-E to support image prompts in addition to text inputs, by taking snapshots of model-objects workspace in their workspace and generating variations of that input. Zhang et al. [81] introduced StoryDrawer a co-creative drawing system to support children in creative storytelling through interacting with an AI through a conversational dialog and drawing. In the current literature, investigations towards more expressive prompting are scarce. However, enabling users to better express themselves when interacting with prompt-based models is crucial to support more complex workflows such as world building. Therefore, we add two new dimensions of expressive prompting to the current literature, namely hierarchical prompting and spatial prompting.

2.3 Scene Generation

Although our focus is on the visual representation of 2D fictional worlds, previous research has already examined text-based worlds [10, 33] and virtual 3D worlds [82]. Anticipating the emergence of language-based 3D scene generation systems, Coyne and Sproat [11] presented *WordsEye*, a tool that automatically converts text into 3D scenes. However, this tool relied on a vast database of pre-existing 3D models and poses. In contrast, fictional world-building typically involves the creation of new artwork. Consequently, this method may face limitations when dealing with unstructured 2D fictional worlds or unconventional layouts. According to a recent survey [80], interactive text-to-scene systems are relatively scarce, while most related work has concentrated on automated approaches [28, 36, 41], neglecting the role of humans in the creative process. Nonetheless, there are some examples of systems that adopt a more human-centric perspective [11, 57]. Our system, WorldSmith, is intended to assist users in creating rather than substituting for them.

There has been growing interest in the development of interactive scene-generation systems. One approach is to use a scene graph to generate images, as explored in [54]. Another interesting direction in interactive image generation is the use of chat interfaces [20, 67]. However, fictional worlds often have a spatial component, but it has been found that human language for expressing spatial relations is often ambiguous and subjective [49, 56]. We designed WorldSmith to allow users to draw their spatial knowledge in addition to text input.

2.4 Human-AI Co-Creativity

Co-creative systems that involve both humans and Artificial Intelligence (AI) entail collaboration where each party contributes their capabilities to the creative process. AI systems can assist with generating ideas [68, 77], provide inspiration [24, 34, 74], and support internal reflection on the content [15, 43], while humans can provide subjective judgment and critical thinking. Recently developed co-creative tools include FashionQ [34] for ideation in fashion design, and WeToon [38] which enables users to generate sketches through direct manipulation of a graphical user interface. Other tools employ prompt-based AI models to support users in their design process [47, 48]. However, Jakesch et al. [32] found that generative model biases can influence users' behavior and lead them to choose the most convenient option, often the first generated content item. This insight further highlights the need for the careful evaluation and design of co-creative systems. To this end, various frameworks have been developed to support the design of a creative partner [1, 50, 61], given the challenges of developing effective AI. Moreover, several works have suggested to also logging users' interactions [16, 17, 42, 48] to evaluate their behavior with co-creative AI systems. With WorldSmith, we contribute a creativity tool to support users' world-building workflow through iterative and expressive interactions.

3 FORMATIVE STUDY

A formative study was carried out to gain insight into the approaches and perceptions creators have in constructing and defining their fictional worlds.

3.1 Participants and Procedure

A total of N=4 individuals participated in remote interviews, with ages between 25-52 years. Participants were recruited through internal email lists and one external professional was recommended by personal contacts. All participants reported prior experience in world-building, either as a hobby or within a professional context designing landscapes or animating game assets. Two participants spent less than an hour per week on world-building tasks, one

spent 1-5 hours per week, and one spent at least 5 hours per week constructing worlds.

During the study, semi-structured interviews were conducted to investigate how participants plan their world-building process. On average, each interview took 45 minutes. The guiding questions were focused on how the study participants characterized world-building and the methods they employed throughout their world-building process.

3.2 World Building Process

From the results of the interviews, we found that participants had different motivations for building worlds as well as different tools used, though many followed similar processes. While some participants (P_1, P_3) found joy in developing interesting and fun worlds to share with their friends, other participants created intricate worlds as part of their profession. For example, P_2 taught about a landscape design class at a university and builds terrain maps to investigate how architectural structures evolve over time based on the surrounding environment. P_4 is a professional game animator who has animated multiple 3D fictional worlds.

3.2.1 Finding Inspiration and Re-using Assets. When it comes to finding inspiration for their creative work, many commented that they often turn to the internet. For example, they reported using search engines or browsing related blogs and online communities like Reddit in search of visual imagery that sparks their imagination (P_1, P_2 P_3). In a Dungeon and Dragons campaign, P_1 had the primary responsibility of constructing the game's world. However, they found creating visual assets tedious and time-consuming. As a result, P_1 frequently resorted to using pre-existing assets found online and noted that there may be copyright issues with re-using these assets.

3.2.2 Refining Ideas. The initial image or inspiration is often vague and often requires many iterations before it becomes something concrete. From their initial spark, participants began asking themselves questions about the fictional world they wanted to build, using an image sketch or list of notes as a starting point to generate further ideas and details. As one participant explained, *"Sometimes a piece of visual artwork strikes me and I find myself asking: What is happening in that image? What is the character doing in that scene?"* (P_4) This iterative process of questioning and building upon ideas aids the development of richer and more detailed fictional worlds.

3.2.3 Inductive Rather Than Divergent World-Building Process. On a macro level, rather than exploring many divergent creations, it is more common to fix key characteristics such as the time and age, or style of the world (fantasy world, science fiction world) once at the beginning of the world building process. Fine details of the world such as individual fauna, flora, and their spatial composition are subject to more frequent changes. However, the further a creator is in the design process, i.e. the more complex the fictional world is, the less likely are major changes in the image composition, because such changes would introduce too much re-work.

3.2.4 Throwaway Prototypes. After the initial idea finding phase, a key activity is creating disposable prototypes quickly. There are many tools to support world builders to accomplish their task, but these tools are highly specialized and often have a steep learning curve to master them fully. Participants in our study reported that they use multiple tools during their world building process. Nevertheless, all participants often started with a rough outline of the world they want to build using pen and paper only, because it allows them to quickly note and depict ideas. These outlines may only include a list of notes (P_3), but more often also include graphical elements such as a mind map (P_4) or layout sketch with text annotations that describe which elements should appear in their artwork later on (P_1, P_2).

3.2.5 High-Fidelity Artwork. After an initial stage of ideation, the desired level of professionalism and the complexity of the fictional world determines which tools they use. As a hobby Dungeon and Dragons game master, P_1 use specialized online world-building tools for generating graphical game elements such as maps [3, 31], assets [26], and characters [55]. Furthermore, P_1 felt overwhelmed with learning all the tools required to create various elements for the DnD campaign. P_3 is an experienced software developer and built a computational agent that procedurally created a DnD map. During the DnD campaign P_3 would use a printed version of the previously generated map and let players spontaneously draw additional game elements on it. Although it saves time, P_3 wanted to have an image for the player to set the "mood" for the current DnD campaign. For landscape design, P_2 use a range of terrain editing software [25, 51, 52, 58] to model photo-realistic landscapes. Here, P_2 noted that, although these tools produce highly realistic 3D terrains, this comes at the cost of a high learning curve.

4 DESIGN GOALS

We formulated the following design objectives, based on the processes identified during our formative user study (Section 3.2), to assist users in constructing their fictional worlds.

D1 - Support Multi-Modal Input World-building includes multiple steps that require prototypes of different fidelity. Early prototypes are usually coarse and are mainly text driven, sometimes also including simple sketches. Our system needs to support multiple input modes to allow users to express their design intent.

D2 - Supporting Iterative Refinement The system needs to allow users to continually incorporate new details into their world. Therefore, the prototype should facilitate the ability to make layered revisions to images that users have previously created.

D3 - Support Visual Asset Generation In order to facilitate the visualization of intricate worlds, the prototype should empower users to create new visual assets to populate their world.

D4- Enable Hierarchical Composition The formative study indicated that world builders typically work on various levels of detail. At the macro level, they establish the general layout of the world, while at the micro level, they determine the specific components present within the world. Hence, our prototype should allow users to engage hierarchically in the design process.

5 WORLDSMITH

We designed our prototype (Figure 1) to support users' world-building workflow by allowing them to focus on different sub-components of their world and perform layered edits. Through multiple generated images users iteratively refine their initially

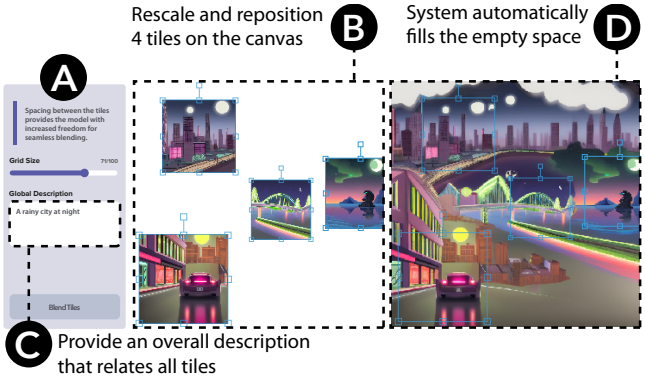

Figure 2: Figure depicts the *Global Tile View*, where all the tiles have already been created by P_{11}B). The tool panel on the left-hand side facilitates the control over space between individual tiles (A) and, further, allows for a description of how the tiles should be blended together (C). The text prompt reads: "a rainy city at night."

vague ideas. An interactive *Tree View* allows them to introspect their past actions and branch out to create new visual assets all in one application.

5.1 Global Tile View

Inspired by the game Carcassone, WorldSmith lets users create a world image with multiple *Image Tiles* (**D4**, Figure 2 (B)). WorldSmith includes four image tiles which gave participant in our user study (Section 6) enough time to work on each tile in detail. However, our concept also allows for more image tiles. Tiles are initially aligned in a grid but can be resized and moved on the canvas.

WorldSmith supports multi-level editing with the *Global Tile View* which allows users to blend multiple tiles together to form a cohesive world (**D4**). We decided to separate the image tile composition from the image tile creation, as they conceptually represent different layers of abstractions. In addition, this division allows users to concentrate on broader objectives, such as combining various tiles that depict the creative vision of the world, while deferring the intricate editing process for each image tile. Typically, these components are guided by a narrative framework; for instance, a user may wish to construct a map featuring a forest, a lake, and mountains. Once blending is complete, the result is shown next to the global tile view, and users can return to individual tiles for further editing if necessary.

Blending Tiles Users can blend tiles by providing a text prompt (Figure 2) and adjusting the empty space between them. The system fills the space between tiles based on the created tiles and a text prompt.

Resizing and Repositioning Tiles The *Grid Size* slider controls the amount of empty space, with more space providing more blending space. Users can also resize and reposition tiles on the canvas for added flexibility.

5.2 Detail Tile Editor

The *Detail Editor View*(Figure 1B) allows users to focus on a specific part of the world within a larger composition. It provides several

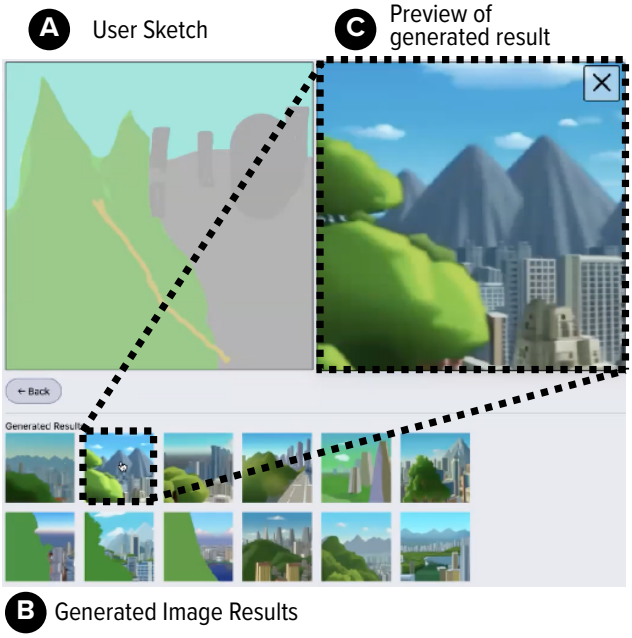

Figure 3: (A) An example of a sketch from P_3 with the overall scene text prompt: "A skyline view of a city in the Caribbean, it has a chain of mountains in the left, and next to them there is a big city, it has a lot of skyscrapers, a path is coming down from the mountain and integrating into the city, Anime style". (B) The generated results are shown at the bottom and (C) enlarged when hovered over.

tools that allow users to generate content using text, sketching, and masking (**D1**).

5.2.1 Text Prompt Editor. The text editor of WorldSmith comprises of two sections: a global scene description area and a specific region description area for users to provide more focused descriptions.

Overall Scene Description The overall scene description text box provides a plain text entry box for users to generate image content quickly. This is a common UI pattern already found in many recent generative image generation tools [53, 60].

Region Description The region description permits users to spatially specify where content should generate on the canvas. Adding a region inserts an empty text segment with a newly assigned color (see Figure 1B). Users can modify this segment by typing a description and only affecting the outlined area. Regions can be drawn to visually link to the corresponding text segment color.

5.2.2 Large Canvas. A large multi-purpose canvas allows users to draw sketches and iterate on their generated images (Figure 4). Users can choose between a sketch mode or a region mode when iterating on an image (**D2**). Both modes support a textual description (see Section 5.2) that instructs the system how to interpret users' spatial inputs.

Sketch Mode The sketch mode lets users draw sketches using a pen tool (see Figure 3 c). WorldSmith takes the sketch input and a textual description to generate images that are similar to the user's drawn images but adds more detail to the generated image (Figure 3). This allows users to coarsely sketch their image while the system generates a higher fidelity image. They can also drag

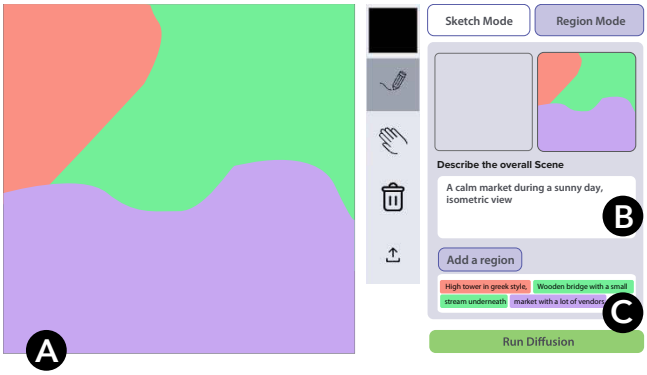

Figure 4: An example of a region segmentation (A) from P_6 with the corresponding scene (B) and region (C) description in the prompt editor. The regions and corresponding text segments are shown with the same color.

existing images and generated images into the sketch canvas to create variations of that image to quickly explore alternative image generations.

Region Mode In the region mode users can draw a region on the canvas by using one of several region brushes: 1) The *Pencil Brush* allows users to draw simple strokes. Each stroke corresponds to one of the text segments previously defined when users created a new region (see Figure 1B), 2) The *Hull Brush* computes the convex hull of all brush strokes combined since the hull brush was selected. This brush allows users to quickly select large regions. 3) The *Lasso Brush* creates a closed path and lets users quickly draw closed shapes. The masks are only visible when the region mode is active (see Figure 4)

5.2.3 Generating Images. After users are done defining their inputs which may include a scene description, a sketch, as well as a few region descriptions, they can generate an image by clicking on the *Run Diffusion* button (Figure 1B).

5.2.4 Results View. The *Results View* is a collection of all images that users have generated displayed as small thumbnails in a scrollable grid. Hovering over a thumbnail displays a larger preview of the image next to the canvas, allowing users to directly compare the differences between the two. When new images are fetched, the *Results View* temporarily greys out to indicate that new content is currently being generated. Once generated, these images will display in the *Results View* so users can reuse it in all their edits.

To insert a selected image into the canvas, users can double-click on the corresponding thumbnail or simply drag the image onto the canvas.

5.3 Tree View

The tree view records all user actions for each tile (see Figure 6). Each node in the tree view includes an image preview and a detailed text description that contains both the scene and region descriptions, representing a snapshot of a tile.

The tree view supports reflective thinking by displaying previous inputs and allowing users to explore the evolution of their design throughout the session [8, 27]. Moreover, it encourages divergent thinking by enabling users to create alternate iterations (**D2**) and explore new visual assets (**D3**). In the realm of text-based

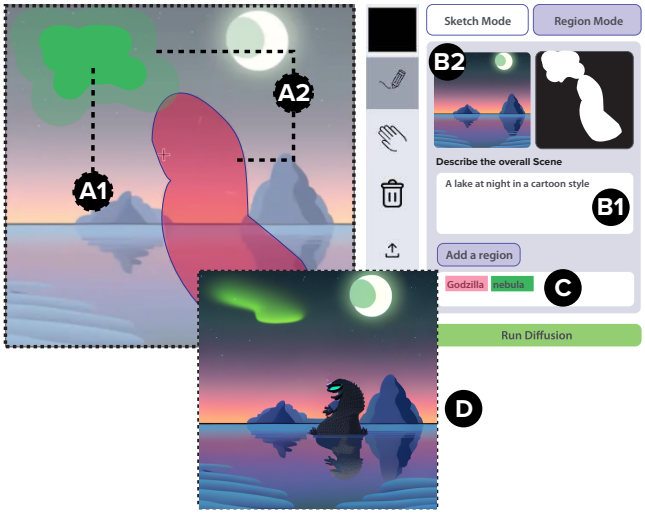

Figure 5: An example of adding sketch information to refine a generated image (B2). P_{11} used a green pen to sketch on the image (A1) and a region tool to highlight parts of the image (A2), thereby providing extra information to guide the generation of a green nebula and Godzilla. The scene description reads: "a lake at night in cartoon style" (B1) and the regions include *Godzilla* and a *nebula* (C). (D) The system redrew the user's input, adding a nebula and Godzilla that matches the style of the image.

world simulation systems, knowledge graphs have been employed to represent the state of the world by mining existing storylines Ammanabrolu et al. [2]. WorldSmith automatically creates such a state tree and additionally enables users to manually extend this state tree.

Our description below outlines the user interactions that facilitated exploring and refining image tiles with the tree view.

Automatic node insertion Whenever users execute the diffusion process, the system automatically updates the tree view. If the inputs to the system have been altered since the last image generation, a new node will be added to the currently selected tree node. This automated insertion upon update enables users to reference their previous interactions.

Furthermore, users have the option to manually insert nodes by selecting a node and clicking on the *Add Node* button. This interaction pattern is useful for experimenting with new-generation ideas while preserving the previous generation. When inserting a new node, users can choose to iterate on an exact copy of the previous inputs, or they can start from scratch to create new visual assets and then blend it into their previously generated world (see Figure 6).

Selecting a Node. Instead of adding a new node, users can also double-click on an existing node in the tree view to load the inputs linked to that particular node into the detail editor view, enabling them to continue iterating on it (**D2**).

Pan and Zoom To accommodate the expanding size of the tree view, we added a pan and zoom function to allow users to zoom in on specific nodes and examine their input. Users can zoom into the canvas using the mouse wheel and pan by clicking and dragging the canvas while holding the left mouse button. Upon generating a

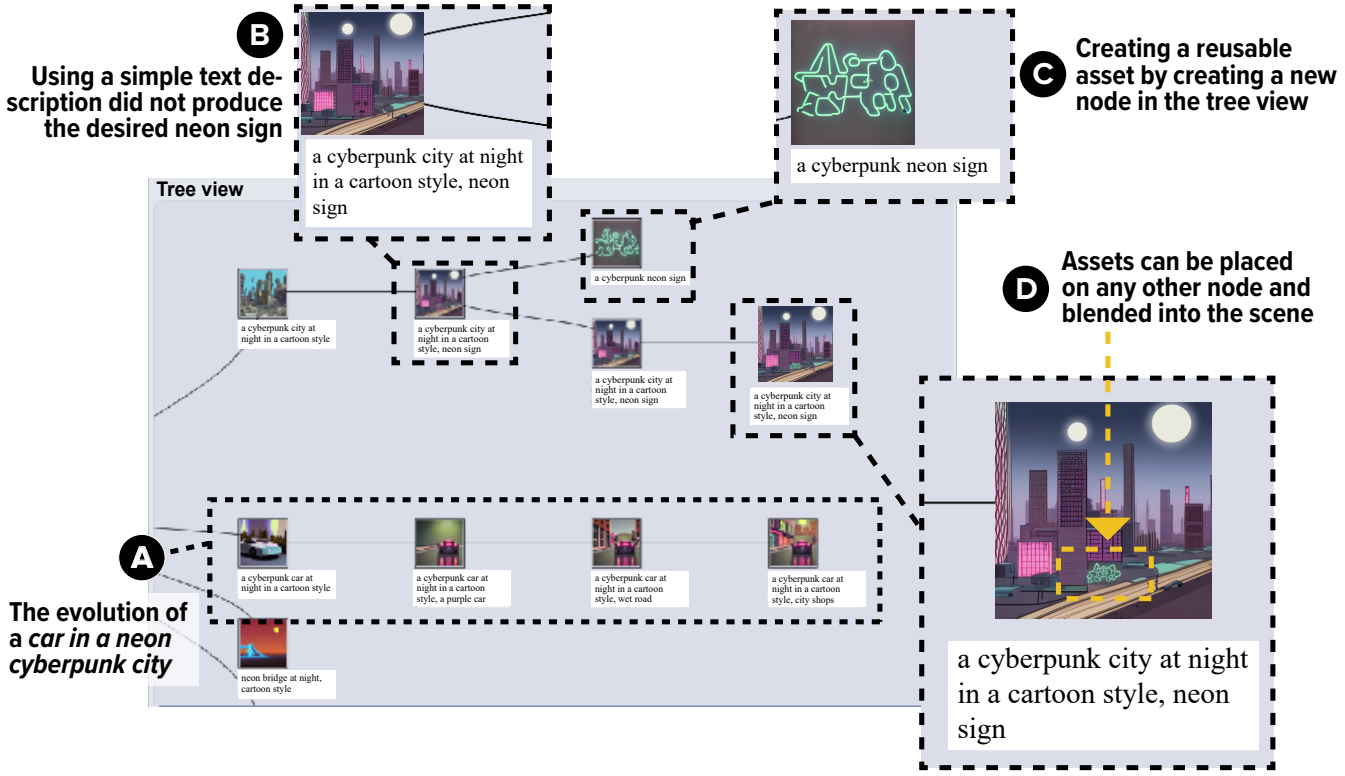

Figure 6: (A) The tree view shows all user inputs and interactions in a tree diagram. This allows users to retrace their creation process, branch out, and create alternative versions of any tile or entirely new assets. (B) During their creation of a cyberpunk city, P_{11} found the neon sign to be lacking. They used the tree view feature to create a neon sign (C) and merged that newly created asset into the previously generated image scene (D).

new node (see Section 5.3), the system automatically centers the viewport to present the node at the center of the screen, aiding users in maintaining focus and navigating the tree view more efficiently.

Preview generated images WorldSmith automatically displays a preview of all images generated for a particular node and its inputs when users hover over an image. This feature allows users to refer to the text prompt beneath the node while also previewing the content they have created.

5.4 Usage Scenario

The following describes one possible scenario of how WorldSmith can be used. This scenario compactly describes how P_{11} in our user study (Section 6) worked with WorldSmith. It demonstrates how the different components of WorldSmith interplay to enable a user to iteratively generate a visual representation of a fictional world, by using text based prompting, iterative workflows, and detailed editing using sketching and regions.

P_{11}, an avid reader of fictional novels and a creative writer, struggles to convey the mood of her imaginative world to her collaborators. To illustrate her world and build a shared understanding with them, she uses WorldSmith, to iteratively create a visual representation (e.g., a map view) of her world from scratch.

To start, she selects one of the four empty image tiles in the Global Tile View (Figure 1 A) and enters "*a lake at night in a cartoon style*" as the scene description (Figure 5 B1). After generating and selecting her favorite initial scene it appears as an input preview (Figure 5 B2). However, it is missing the all-important Godzilla. She switches to the Region Mode (Figure 5 A2, red) and enters the text description "godzilla" as a region descriptor (Figure 5 C). She then adds a Nebula using the Sketch Mode + Region Mode to draw a green mist on the canvas (Figure 5 A1+A2), which the system repaints, blending it in with the rest of the scene.

For the second tile, she starts by generating a scene with overall description "*a cyberpunk city at night*". She then wants to add a reusable neon billboard sign asset. To do that, she adds a new node in the Tree View (Figure 6), clears the canvas, and creates a "*neon billboard*". Once happy with the result, she navigates back to the previous city scene using the Tree View and drags and drops the neon sign into the scene. She masks out the borders of the new image asset, provides a short description of how the asset fits the scene, and the system blends the images.

Using the same techniques, she creates another tile showing a *neon bridge at night* and a *cyberpunk car at night*. P_{11} merges all four image tiles into one coherent image using the Global Tile View (Figure 2 A), adjusting the grid space and provides a short description that relates all tiles (Figure 2C). She positions and resizes (Figure 2B) the tiles to her liking and clicks Blend Tiles. The system contextually fills in the space between the tiles (Figure 2D, Figure 12), and after some iterations.

5.5 Technical Details

Our prototype uses a client-server architecture. The client frontend is built with SvelteKit [70] and Skeleton UI toolkit [69], while the backend is built with FastAPI and runs on the python webserver uvicorn. We utilized the Stable Diffusion algorithm via HuggingFace's model hub (512x512 pixels), which were run on a machine with 16GB GPU VRAM. To enable interactive painting, we used fabricJs[21] and created two separate canvases for sketching and region masks. When users initiated the diffusion process, we extracted inputs from the detail editor view. For scene descriptions without additional input (e.g. sketches or regions) we generated an image from random noise based on the provided text input. If users provided an RGB sketch, we used it with added Gaussian noise to generate images that matches user's drawings following Rombach et al. [63]. Regions were transmitted as an array, with each entry containing a binary mask image and corresponding description. In our region-based painting feature (Figure 4), we extract multiple binary masks from a user-provided region segmentation, where white pixels correspond to the unique region color, and the remaining area is black. Our region-based painting feature is inspired by Balaji et al. [4] who proposed an approach allowing users to specify where elements should appear on the generated image. They combined a separate binary image mask, with dimensions matching the output image, with each word in an input text prompt. Notably, words in the user-provided text input exert variable influence on different parts of the image, with white pixels serving as indicators for a higher probability of an element appearing in the assigned segment. We used an open-source implementation of this concept applied to Stable Diffusion[1]. For blending image tiles, we obtained a binary mask with black pixels for image tiles and white pixels for *empty* space. A Gaussian blur was applied to the mask, softening black tile edges for a smoother blend. The final image was generated by inputting this mask and user-created image tiles into the diffusion model.

To track user interactions (such as typing a scene description, drawing a sketch or a region, moving the tiles), we implemented a logging server on FastAPI, which utilized a Postgres database.

6 EVALUATION

To evaluate the utility and use of WorldSmith, we conducted a user study with 13 participants which provide first insights into the following research questions:

- (*RQ*1) How do users engage with generative AI for world building?
- (*RQ*2) To what extent does WorldSmith support the world building process?

6.1 World Building Task

A set of prompts were created to inspire participants to build their fictional world. The prompts were intended to cover various types of visual world-building, such as fantasy maps or fictional landscapes. The prompts were open-ended and designed to allow participants to think creatively and explore a breadth of concepts (Table A.1 in the Appendix).

[1]https://github.com/cloneofsimo/paint-with-words-sd

We further encouraged them to explore not only maps but also other types of visual imagery depicting fictional worlds. While maps typically have specific characteristics such as a top-down view and a specific scale, fictional worlds can be composed of any objects that create a scene that does not exist in reality. We instructed participants to read the previous design prompts and focus on the composition of elements and the setting of the world (e.g., a fantasy world and everything such worlds entail).

6.2 Methods

We conducted a first-use study online via Zoom and asked participants to think aloud during the user study to learn about their motivations and understand their potential challenges when interacting with the prototype. Our study involved a world-building task where participants used WorldSmith to create a fictional world (Section 5). In addition to logging users' interactions, we conducted semi-structured interviews at the beginning and end of the world-building task to collect their open feedback about their prior experience and motivations for world-building and their overall experience working with WorldSmith. Finally, participants completed two questionnaires at the end of the task where they rated their experience with the different features of WorldSmith. For qualitative analysis, two researchers independently assigned inductive codes for a subset of the transcribed interviews. Then one researcher used these codes and notes taken during the interviews and thoroughly reviewed the full transcripts of all interviews to find further evidence for the thematic clusters identified in the previous step.

6.3 Participants and Procedure

We recruited the participants for the study over e-mail lists and personal contacts. All participants had experience building worlds (between 2 and 10 years). This included experience crafting DnD worlds, video game levels, creative writing, and landscape design. This study was approved through our institutional review process.

The study consisted of four phases which we briefly outline below. Before the study, all participants completed a consent form and demographic questionnaire.

Pre-Interview (10 minutes) The first phase consisted of a short semi-structured interview with each participant. During this interview, we asked participants about their motivations and experience in building fictional worlds.

Tutorial (10 minutes) During the tutorial phase, participants were introduced to the prototype and given an overview of its features and functionalities. One researcher explained the process of creating an example image tile using text, sketching or region painting. Meanwhile, the participants engaged with the tool by following the researcher's guiding instructions to become acquainted with the different ways of providing input.

World Building Task (60 minutes) In the third phase, participants interacted with the software prototype via screen sharing and remote control. They were given a fictional world-building task to complete (Section 6.1). We also asked them to briefly describe what they wanted to create and we recorded their interactions with the software and resultant words.

Post-Interview and Questionnaire (10 minutes) After completing the world-building task, we conducted an interview with each

participant to collect their feedback and impressions of the software prototype. In the questionnaire, we asked participants to rate the features of the prototype on a Likert scale. This questionnaire included questions about the various input modalities (text-only, text+region, text+sketching) and whether they think WorldSmith can speed up their regular approach to world-building. Finally, we administered a *System Usability Scale* (SUS) questionnaire to evaluate the overall usability of WorldSmith which included questions about the complexity of the program, and related how easy it was to learn the program and to interact with it.

6.4 Quantitative Findings

Overall, we found participants leveraged all forms of interactions. In total, participants triggered the image generation process 229 times to generate scenes, maps, and assets for the individual image tiles, resulting in 2748 generated images. Additionally, users created 86 world compositions using the blending features.

In our first-use study, 13 participants interacted with the prototype. However, some participants finished their world-building task before the official time ended, so they started a second session within the remaining time. Therefore, for the remainder of this section, we denoted each of the resulting 16 sessions with a participant id and the corresponding trial number (P_X $Trial$ Y).

6.4.1 Relational vs. Quantifying Keywords. We counted the number of words in 1) the *text of each scene description* and 2) the *text of each region description*. We found that participants wrote differently for scene and region descriptions. Each participant created on average 2.05 regions with corresponding region descriptions per tile (Median=2, inter-quartile range 2). Region descriptions were typically shorter (Avg=4.2 words; Md=3; iqr=6) than scene descriptions (Avg=12.4 words; Md=11; iqr=7). Scene descriptions were primarily used to describe multiple objects and the overall style, while region descriptions were used to quantify or describe a specific object or element and to provide more localized instructions.

Position – We analyzed scene and region descriptions using a coding routine similar to that in Section 6.2, using text prompts from user interaction logging data. The full list of codes can be found in Table A.2 in the Appendix. Our analysis found that scene descriptions contained more *positional* keywords (e.g. *surrounded by, above, north, south*) than region descriptions, which is consistent with participant observed behavior in the user study where they used the region drawing tool to specify spatial relations. This is shown in Figure 7.

Style – Scene descriptions also more frequently included *style* (e.g. cartoon style, fantasy) and *perspective* keywords (top-down view, isometric view). Participants used the UI in overall scene descriptions to define *style* and *perspective* keywords rather than repeating them for every region description. More generally, participants wanted to define these keywords once for the entire session and for all image tiles.

Action – Scene descriptions frequently included *action* keywords, which relate one object to another (e.g. "Mountain range *running* north to south"). In contrast, region descriptions typically described only one object without trying to relate it to other elements in the world, which is done implicitly via drawing the regions.

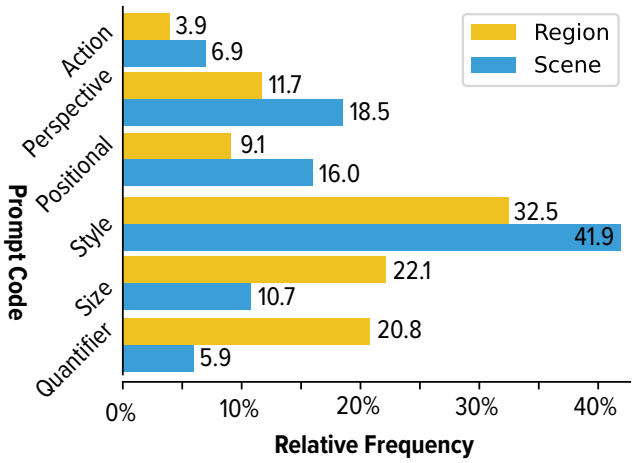

Figure 7: An overview of the codes extracted from the prompts for region descriptions and scene descriptions. See Table A.2 in the Appendix for an explanation of the codes. The blue and orange bars add up to one, respectively. Note that scene descriptions frequently contained style and perspective-related keywords, while region descriptions contained more size and quantifier keywords.

Quantifier – Participants in our study used indefinite quantifiers (e.g. "a few trees", "many dirt roads") and size keywords (e.g. "a large", "a gigantic") more frequently in region descriptions to convey the number and size of elements in the world, instead of drawing separate regions for each.

6.4.2 Composing Tiles. The final iteration with WorldSmith often involved rearranging and resizing the created image tiles (Figure A.1). The prompts at the global level were more generalized compared to the individual tile level prompts. They included keywords to remove visual artifacts such as border around the tiles during blending (*"no [straight] edges, smooth collage"* (P_2), *"seamless map"* (P_1)) or introduce new objects to connect the tiles (*"[...] islands connected by bridges"* (P_5), *"[...] an ocean in the center"* (P_8)). Although, keywords related to style and perspective were also used, global keywords didn't affect already created image tiles since WorldSmith does not support perspective and style matching across tiles post image tile creation.

6.4.3 Interaction traces. Users' editing behavior was analyzed by filtering four representative actions from the interaction logging data (Figure 8). These actions include modifying tiles, sketches, regions, and text prompts. Examples of modifying tiles include repositioning and scaling them in the *Global Tile View* (Figure 1 (A)) while modifying sketches involves drawing on the Sketch Canvas. Modifying regions involves adding, drawing, and describing new regions, and modifying text prompts only involves editing the scene description in the *Detail Editor View*.

Bootstrapping the World with Text – Participants in 12 out of 16 sessions began creating their world with a text-only description, likely due to the fast and lightweight nature of the method, as reported by participants in the study: *"I do like just having the general big description, and just seeing what it comes up with."* (P_{11}). Participants in the other 4 sessions started with an initial sketch of

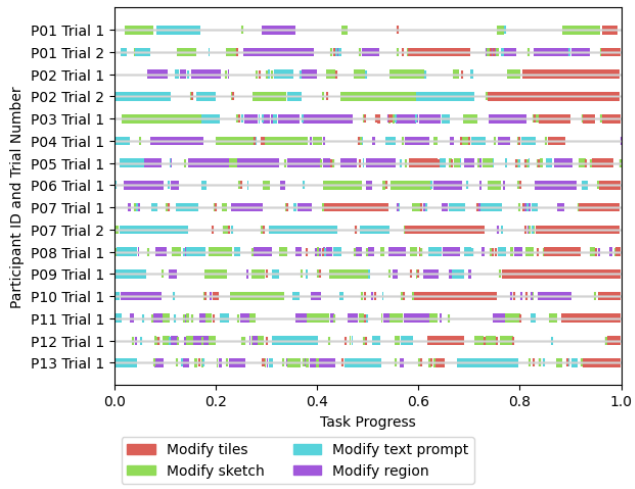

Figure 8: The graph shows the distribution of participants' interaction traces while they worked on the interactive prototype. Three participants completed the task early and started a second world-building session, resulting in a total of 16 sessions instead of 13. Note that most participants started with a text prompt when first interacting with WorldSmith.

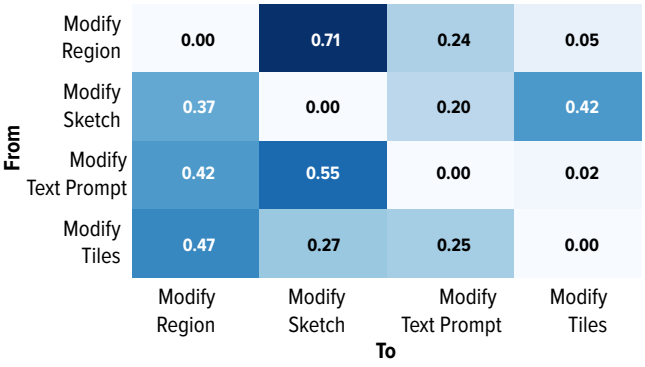

Figure 9: An overview of the action transition frequencies aggregates across all sessions. The number in each matrix cell represents the relative transition ratio at which a participant transitioned from action Y (rows) to action X (columns).

their worlds. These participants already had an image with a rough structure in mind. Using the sketching tools directly, they could sketch out their mental image.

Moving from Coarse to Detail – We computed and aggregated the transition matrices between creation operations over all participants and trials to analyze editing behavior (Figure 9). Overall, participants have transitioned between all available edit operations. However, modifying regions frequently preceded the blending process the last action in the world-building process. This observation, together with our previous observations (Section 6.4.3) suggest that participants transitioned from coarse actions, such as text prompts, to fine-grained editings, such as sketching and region painting, to add details to their image tiles. This was consistent with the previous observation that participants used text prompts to quickly bootstrap the world-building process.

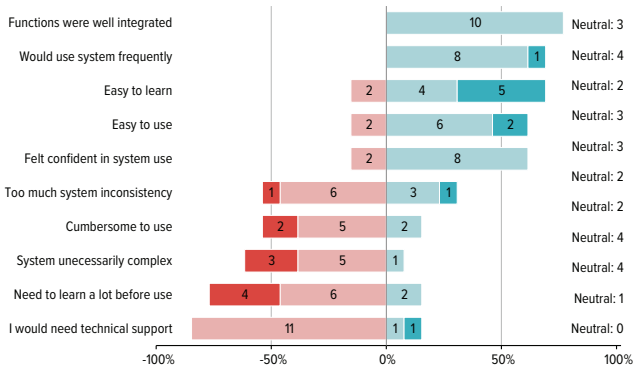

Figure 10: An overview of the responses for the System Usability Scale questionnaire.

Summary – The quantitative results indicate that users engaged with WorldSmith in different ways (*RQ*1). They generally moved from coarse to detail, i.e. using text and regions to populate the individual tiles before sketching in the details. On the global composition level, participants explored alternative worlds by revising both their prompts as well as tile compositions. Differences in language used in their prompts imply that users perceive scene and region prompts differently and convey input information implicitly through both text and sketching.

6.5 Design Insights and Improvements

Overall, the participants felt that the tool was easy to use and helped them perform the world-building task (Figure 10). Our observations and participants' comments during the interview revealed that some system features required further iteration.

Related Keywords – As noted by Liu and Chilton [46] and Liu et al. [48], developing relevant keywords for text-to-image generations is challenging. While participants in the study were primarily concerned with building their fictional world, they felt burdened to think of keywords that are relevant to their world, such as stylistic keywords and perspective. They wished the system would automatically insert those keywords and match the style across all tiles.

Direct Manipulation – There was a desire to more directly edit the result image, with one participant explaining *"It would have been nice to be able to manipulate elements in the generated image directly. For example, after I created that tile with a crossroad, I wanted to select that road as a spline object that could be directly manipulated."* (P_2). This would allow for greater control and precision in the image creation process, though comes with a large set of technical challenges.

6.6 Qualitative Findings

6.6.1 Working with Region-Based Descriptions. We observed two distinct strategies when participants interacted with the region-based descriptions.

Narrative first, then drawing – When defining regions, some participants created the text segments first and typed the full description of each region before deciding how these regions are composed on the canvas. This is in line with findings from the

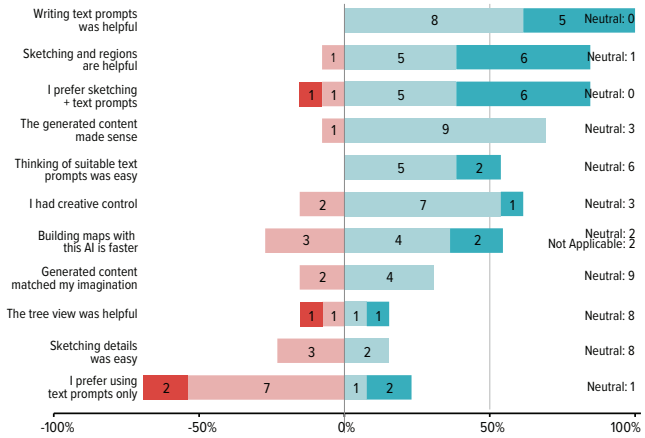

Figure 11: An overview of the responses from the final questionnaire.

formative study (Section 3.2) on starting with a concept of the fictional world first before committing to a specific instantiation of that world.

Drawing first, then narrative – On the other hand, some participants first thought about the elements' overall composition before writing down a detailed description for each element. When asked whether they had a concrete image in mind, they replied with: *"I don't have anything specific in mind, but rather a rough idea of the elements I want in that picture."* (P_{13})

6.6.2 Users preferred Multi-Modal Input. While all participants agreed that writing text prompts to generate images was helpful, the majority (11 out of 13) preferred multi-modal input over using text-prompts only (Figure 11). Deriving suitable text prompts for image generation was not considered difficult, however, we observed during the study that participants valued content that described the core-elements of their fictional-world (e.g. a magic forest, a mage rabbit), more than keywords which related to overall rendering features such as stylistic and perspective keywords. Participants also responded that they preferred adding sketches and painting in addition to their text input. From our observations we note that some participants (P_2, P_3) found it challenging to sketch-in details with the remote controlled computer using the mouse, while other participants saw this technical constraint as a strong motivator for generating images from rough sketches without the need for precise 2D input (P_6) using WorldSmith.

6.6.3 Building World by Parts. During the formative study, we found that not all parts of an image hold equal importance for the creator. This is particularly evident in Dungeons and Dragons (DnD) games, where certain areas of interest contain more detail and are the sites of important events. Observations from our first-use study underpinned the previous finding by showing that participants invested significant effort in creating highly detailed individual tiles, and appreciated the ability to focus on these tiles while leaving the empty spaces between them to be filled in automatically by WorldSmith to smoothly combine the tiles. Additionally, one participant noted that *"[t]he tiles have lots of detail which naturally draws the viewers eye to those points of interest."* (P_1). Another participant found blending to be useful to explore different world compositions quickly: *" You know, sometimes [...] I know that I want there to be like a lake over here in a city over here. But I don't really care what else is in there [...]. Let's throw together some interesting stuff and see how it blends together, and then start adding on from here."* (P_4).

6.6.4 Materialise Stream of Consciousness. Participants commented that they sometimes struggled to track their thoughts because *"ideas flash up and vanish"* (P_{13}) before they had the chance to fully develop that idea. However, using image generation has aided them in quickly capturing their train of thought. One participant further noted *"I'm one of those people who has a very blind inner eye, so I can't visualize things in my head. I have to put it in front of me in order to make any sense of it visually."* (P_{11}). Using the multi-modal image generation system has enabled them to refine their initially *vague ideas* by allowing them to continuously add details to their creations. Here, one participant noted that *"seeing how all seamlessly blend, I now had a clearer vision of how I want to compose the elements in the world"* (P_3).

6.6.5 Perception of control over the AI. Participants were generally conscious that they were interacting with an AI system. During the world-building task, we frequently observed participants questioning whether the system understood their commands or intentions. As a result, participants appeared to be more understanding when WorldSmith failed to generate exactly what they had in minds. In such instances, participants began to consider alternative ways of expressing their intent. They sometimes reformulated their text input to the system or included supplementary information for the system, such as drawing a sketch. For example, one participant wanted to create a nebula on a night sky using the region-based painting tool, but the system did not produce the desired result. In response, the participant decided to add more information to the input by also adding a sketch in addition to the region-based painting (see Figure 5). However, some participants were discouraged by their initial failed attempts to author their envisioned image tile, feeling that the system was inconsistent (Figure 10) when generating new images (P_4, P_7, P_8). For example, while P_4 could use WorldSmith to generate *"ducks in a pond"* or *"a house surrounded by a forest"*, he was struggling to generate a complex scene such as *"a top-down view of a mountain rift running North to South"* and wanted to insert a *"a fantasy art of a mouth of a mine with mine carts arranged around"* and a *"top down view of a yard surrounded by tents and roman soldiers"*. Creating such a scene would have required more time to create relevant assets such as the mine or the war camp.

6.6.6 Feedback on the Tree View. Overall, we found that those participants who used the tree view explored this concept to create new image assets and blend them back into their scenes. One participant mainly used this feature to introspect his past interactions. The tree view enabled him to be more confident in exploring alternative generations because it offered a way to revert to the original image. *"I really liked the tree view, because I use a lot of [...] platforms and I'm often hesitant to continue iterating on an idea [...] because it's hard to get back to the original image. Sometimes it kind of gets lost, even if it's somewhere in the history it's up to you to find the original image that you originated from [...]"* (P_{10}). Another participant commented that she wanted to create *"unlimited nodes and add them directly as*

image tiles to the global canvas" (P_{12}). However, the responses in the final questionnaire (Figure 11) indicated that participants were generally neutral about the *Tree View* in this study. We reflect on this observation in the discussion.

6.6.7 Feedback on the Generated Worlds. Participants in our study have found that the generated worlds made sense, although they did not always match their prior imagination. This included the generated individual tiles as well as the overall blended tiles (Figure 12)

During the interview 11 out of 13 participants commented positively about the blended results, highlighting that the system was capable of sensibly filling in the gaps between the created image tiles. The other two participants generally liked the idea of blending different parts of their image but found it challenging to create a seamless blend of all their tiles.

6.6.8 Comparisons against world-building without Generative AI. When participants reflected on their experience with WorldSmith, two participants commended the fast generation of an initial map draft by focusing only on a few *key areas* (P_1,P_9). Nevertheless, full utility of the generated maps required better style and perspective alignment across all tiles (P_6, P_9, P_8). For example, P_8 in found it difficult to align a tile with *"a cartographic rendering of an arctic tundra"* with a *"cartoon style desert island with war camps"*.
Aside from DnD maps, one participant particularly liked the ability to customize tiles. He compared the tile generation process using WorldSmith and his prior experience with in-game map builders noting: *"In the past, while working with tile-based map editors, I frequently found myself searching for compatible tiles, for instance, connecting streets. However, with this system, I would simply provide sufficient space between the tiles and let the system determine the best way to merge them"* (P_6). P_{11} in particular found that WorldSmith's proposed workflow closely matches her own world-building process: *"I do the more fine details and then go. Oh, wait! I should do something [...] a little bit more broad. So for me [the workflow] wasn't anything new. It was kind of a more natural flow for me."*

Summary – In relation to ($RQ2$), our findings demonstrate that the suggested workflow of WorldSmith effectively complemented participants' existing world building process, enabling them to swiftly produce a rendition of their envisioned worlds. During a 45-minute user study, each participant successfully generated a version of their world. While WorldSmith facilitated the rapid creation of an initial draft, participants also acknowledged the need to address appropriate scaling, stylization, and perspective coherence among the already generated image tiles in order to accurately represent their envisioned world.

7 DISCUSSION

7.1 Speed vs. Quality Trade-Off

We observed that the quality of the blended results varied among participants. Some participants could seamlessly blend multiple tiles, while others required multiple iterations. We observed three main factors that influenced the quality of the blended result in our study: 1) Tiles with similar styles were easier to blend. 2) Participants had to consider the logical structure of tile composition, such as creating a single horizon across multiple tiles to achieve a seamless blend. Ambiguous positioning of the horizon could result in a less seamless blend. 3) Tile complexity, including the length of the text prompt, number of regions, and style parameters, also affected blending quality. More detailed tiles were harder to blend seamlessly.

Figure 12: Figure shows multiple worlds that users have built using WorldSmith. Some show a fictional map, while others depict a scene in a fictional world.

We experimented with MultiDiffusion [5] and found it produced better results than stable diffusion but was significantly slower (over 30 seconds for a 768x768 pixel image). Therefore, we prioritized speed and used stable diffusion for image generation to allow users to iterate quickly. Our main objective was to analyze how users developed their fictional world. Some participants were satisfied if the approximate content of their initially created tiles was preserved for the final blending. They were also willing to wait for the final rendering once they were satisfied with the overall composition. To address this, future work could include a slower, quality-preserving mechanism like MultiDiffusion to be used as a post-processing step for the final rendering.

7.2 Creative Use of Model Bias

We have found that stable diffusion [63] excels in generating images where the elements fit naturally into the scene, such as a spider in a forest, but often faces challenges when generating images of concepts that do not belong together using only text prompts, such as a snow-covered mountain ridge on a tropical island. It has been found that LLMs, sometimes "hallucinate" facts and struggle to generate coherent storylines [14]. However, our users' feedback indicated that visual hallucinations [6] can be a desirable attribute that allows users to create unconventional worlds. Therefore, our prototype offers two possible interactions to combine different concepts. The first solution involves merging tiles that portray

disparate concepts, while the second allows users to blend arbitrary images into the scene using a region mask.

Related work has improved the synthesis of images with disparate concepts [4, 22]. With WorldSmith we complement these systems by providing a tool that enables users to interactively define blending on three levels: 1) blending using text-only descriptions, 2) blending by also providing sketching and region-masks for disparate concepts, and 3) blending of multiple image tiles.

7.3 Limitations and Reflections on Methodology

Our study was conducted with a pool of participants who shared a common interest and proficiency in technical systems. Many had prior exposure to generative image systems, which allowed them to engage easily with the system during the study. We note that participants without prior experience in image editing may require additional time to familiarize themselves with the system. This learning curve may vary depending on the individual's technical knowledge and experience level. The user study suggested that participants mainly required help to find domain-specific language related to world-building, e.g. perspective, style. Nonetheless, we ensured that all participants received guidance and suggested relevant keywords based on the verbal descriptions of what they wanted to create. Future systems can include a LLM to suggest relevant keywords to lower the threshold for using the system [48].

During the study, we found that participants were focusing more on building the individual image tiles, which occupied most of their time. Given the time limit of the user study, few participants interacted extensively with the *Tree View*. However, those who did comment positively about it. We believe that such a tool is best explored over a longer period and over distinct sessions to enable users to introspect their past behavior and explore a wider variety of scenes as they create their world.

In this study we focused on the multi-modal input techniques to leverage a generative AI for world-building. However, we did not consider lore, musical score or character building which are also an essential component of successful worlds.

Our study identified certain limitations with respect to the quality and speed of WorldSmith. We found that the generated images sometimes did not accurately capture all the concepts mentioned in users' text descriptions. Additionally, users had to wait for 6-8 seconds before the first batch of images was generated. While we expect that the development of pre-trained generative AI models will continue to improve the accuracy and speed of image generation, it is important to acknowledge that human language remains inherently ambiguous. As such, we need to explore additional methods to help users express their creative vision when working with generative AI models.

7.4 Expressive Prompting

Current user interfaces for prompt-based models [53, 60] tend to promote a one-shot image generation approach, wherein users can only modify the text to influence the generation process. However, our user study highlighted a key insight: participants tend to create their world models in parts, using sketching and text inputs to communicate positional information. To support this design process, WorldSmith offers users the ability to sketch and paint regions in addition to entering text prompts. Furthermore, WorldSmith incorporates hierarchical support by separating tile blending from tile creation.

Based on our improved understanding of how users integrate multi-modal input in the process of building worlds, we have identified two distinct dimensions that demonstrate how WorldSmith enables more expressive prompting beyond the limitations of the "click-once" interaction paradigm.

Hierarchical Prompting – WorldSmith facilitates hierarchical prompting by enabling users to define prompts across three levels. Firstly, at the Image Tile Canvas level, users can provide text prompts to create a base scene, thus setting the stage for their world-building process. Secondly, at the Sketch and Regions level, users can add text prompts to sketches and regions to provide structural guidance to the system. Finally, users can add text prompts to blend multiple tiles together at the Global Tiles level.

Spatial Prompting – In addition to its other features, WorldSmith also allows participants in our study to modify prompts spatially through non-textual interactions. This capability is facilitated through two methods: firstly, users can convey spatial prompt information through sketching, thus engaging in what we term "Prompting through Painting." Secondly, users can move image tiles to convey prompt information, which we call "Prompting through Dragging."

In summary, WorldSmith introduces two distinct dimensions that enhance users' ability to interact more expressively with prompt-based models. However, we believe that expressive prompting is not limited to world-building, but can also aid the composition of any complex image. Traditional photo editing includes a layering system to allow users to organize and structure their images manually. Spatial and hierarchical prompting augments this interaction and allows participants to quickly explore blended image compositions.

8 CONCLUSION

In summary, we have looked beyond the "click-once" image generation interaction paradigm for prompting generative AI, and we discovered through a first-use study with WorldSmith that users leveraged all inputs (text, sketch, and region masks) in combination. Crucially, participants expressed their creative vision not only through textual descriptions but also through non-textual interactions with the system. Based on our findings, we propose two expressive prompting concepts as part of WorldSmith's graphical UI, supporting: 1) *hierarchical prompting*, which involves the use of layered prompts, and 2) *spatial prompting*, which allows users to spatially arrange prompts. With WorldSmith, we illustrate how these prompting concepts aid the fictional world-building process. Beyond this use case, we see expressive prompting as a general concept to inspire user interfaces that support users' complex workflows with prompt-based AI models.

ACKNOWLEDGMENTS

We thank Justin Maltejka, Jo Vermeulen, Bon Adriel Aseniero, David Ledo, Qian Zhou and Daniel Buschek for providing valuable feedback on this work.

REFERENCES

[1] Saleema Amershi, Daniel S. Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi T. Iqbal, Paul N. Bennett, Kori Inkpen Quinn, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019).

[2] Prithviraj Ammanabrolu, Wesley Cheung, Dan Tu, William Broniec, and Mark O. Riedl. 2020. Bringing Stories Alive: Generating Interactive Fiction Worlds. In *Artificial Intelligence and Interactive Digital Entertainment Conference*.

[3] World Anvil. 2023. World Anvil. https://www.worldanvil.com.

[4] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. 2023. eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. arXiv:2211.01324 [cs.CV]

[5] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. 2023. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. *arXiv preprint arXiv:2302.08113* (2023).

[6] Leonid Berov and Kai-Uwe Kühnberger. 2016. Visual Hallucination For Computational Creation. In *International Conference on Innovative Computing and Cloud Computing*.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[8] Brian Burg, Richard Bailey, Amy J. Ko, and Michael D. Ernst. 2013. Interactive record/replay for web application debugging. *Proceedings of the 26th annual ACM symposium on User interface software and technology* (2013).

[9] Lydia B. Chilton, S. Petridis, and Maneesh Agrawala. 2019. VisiBlends: A Flexible Workflow for Visual Blends. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019).

[10] Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben A. Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew J. Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. 2018. TextWorld: A Learning Environment for Text-based Games. In *CGW@IJCAI*.

[11] Robert Coyne and Richard Sproat. 2001. WordsEye: an automatic text-to-scene conversion system. *Proceedings of the 28th annual conference on Computer graphics and interactive techniques* (2001).

[12] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. 2003. Object removal by exemplar-based inpainting. *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.* 2 (2003), II–II.

[13] João Miguel Cunha, João Gonçalves, Pedro Martins, P. Machado, and Amílcar Cardoso. 2017. A Pig, an Angel and a Cactus Walk Into a Blender: A Descriptive Approach to Visual Blending. *ArXiv* abs/1706.09076 (2017).

[14] Robert Dale. 2021. GPT-3: What's it good for? *Natural Language Engineering* 27 (01 2021), 113–118. https://doi.org/10.1017/S1351324920000601

[15] Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. 2022. Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries. *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (2022).

[16] Hai Dang, Sven Goller, Florian Lehmann, and Daniel Buschek. 2023. Choice Over Control: How Users Write with Large Language Models using Diegetic and Non-Diegetic Prompting. *ArXiv* abs/2303.03199 (2023).

[17] Hai Dang, Lukas Mecke, and Daniel Buschek. 2022. GANSlider: How Users Control Generative Models for Images using Multiple Sliders with and without Feedforward Information. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022).

[18] Pérez Domínguez and Enrique Alejandro. 2019. The design of indie games, a different paradigm.

[19] Stefan Ekman. 2019. Vitruvius , Critics , and the Architecture of Worlds : Extra-Narratival Material and Critical World-Building.

[20] Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, R. Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W. Taylor. 2018. Keep Drawing It: Iterative language-based image generation and editing. *ArXiv* abs/1811.09845 (2018).

[21] fabricjs. 2023. *fabricjs*. https://fabricjs.com

[22] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors. *ArXiv* abs/2203.13131 (2022).

[23] Songwei Ge and Devi Parikh. 2021. Visual Conceptual Blending with Large-scale Language and Vision Models. *ArXiv* abs/2106.14127 (2021).

[24] K. Gero, Vivian Liu, and Lydia B. Chilton. 2021. Sparks: Inspiration for Science Writing using Language Models. *Designing Interactive Systems Conference* (2021).

[25] BiteTheBytes GmbH. 2023. BiteTheBytes World Creator. https://www.world-creator.com.

[26] Dungeonfog GmbH. 2023. Dungeonfog. https://www.dungeonfog.com.

[27] Tovi Grossman, Justin Matejka, and George W. Fitzmaurice. 2010. Chronicle: capture, exploration, and playback of document workflow histories. *Proceedings of the 23nd annual ACM symposium on User interface software and technology* (2010).

[28] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. 2018. Imagine This! Scripts to Compositions to Videos. *ArXiv* abs/1804.03608 (2018).

[29] Gary Gygax and David Cook. 1989. *The Dungeon Master Guide, No. 2100, 2nd Edition (Advanced Dungeons and Dragons).* TSR, Inc. https://www.amazon.com/Dungeon-Master-Advanced-Dungeons-Dragons/dp/0880387297?SubscriptionId=AKIAIOBINVZYXZQZ2U3A&tag=chimbori05-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0880387297

[30] Adi Haviv, Jonathan Berant, and Amir Globerson. 2021. BERTese: Learning to Speak to BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 3618–3623. https://doi.org/10.18653/v1/2021.eacl-main.316

[31] Inkarnate. 2023. Inkarnate Creator. https://inkarnate.com/cookies/.

[32] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. *ArXiv* abs/2302.00560 (2023).

[33] Peter Alexander Jansen. 2021. A Systematic Survey of Text Worlds as Embodied Natural Language Environments. *ArXiv* abs/2107.04132 (2021).

[34] Youngseung Jeon, Seungwan Jin, Patrick C. Shih, and Kyungsik Han. 2021. FashionQ: An AI-Driven Creativity Support Tool for Facilitating Ideation in Fashion Design. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021).

[35] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics* 8 (Dec. 2020), 423–438. https://doi.org/10.1162/tacl_a_00324

[36] Dhiraj Joshi, James Ze Wang, and Jia Li. 2006. The Story Picturing Engine—a system for automatic text illustration. *ACM Trans. Multim. Comput. Commun. Appl.* 2 (2006), 68–89.

[37] Pegah Karimi, Mary Lou Maher, Kazjon Grace, and Nicholas M. Davis. 2019. A computational model for visual conceptual blends. *IBM J. Res. Dev.* 63 (2019), 5:1–5:10.

[38] Hyung-Kwon Ko, Subin An, Gwanmo Park, Seungkwon Kim, Daesik Kim, Bo Hyoung Kim, Jaemin Jo, and Jinwook Seo. 2022. We-toon: A Communication Support System between Writers and Artists in Collaborative Webtoon Sketch Revision. *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (2022).

[39] Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. 2022. Large-scale Text-to-Image Generation Models for Visual Artists' Creative Works. *Proceedings of the 28th International Conference on Intelligent User Interfaces* (2022).

[40] Chinmay Kulkarni, Stefania Druga, Minsuk Chang, Alex Fiannaca, Carrie J. Cai, and Michael Terry. 2023. A Word is Worth a Thousand Pictures: Prompts as AI Design Material.

[41] Bongshin Lee, Rubaiat Habib Kazi, and Greg Smith. 2013. SketchStory: Telling More Engaging Stories with Data through Freeform Sketching. *IEEE Transactions on Visualization and Computer Graphics* 19 (2013), 2416–2425.

[42] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022).

[43] Chenchen Liu, Jierui Hou, Yun fang Tu, Youmei Wang, and Gwo jen Hwang. 2021. Incorporating a reflective thinking promoting mechanism into artificial intelligence-supported English writing environments. *Interactive Learning Environments* (2021).

[44] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What Makes Good In-Context Examples for GPT-3? http://arxiv.org/abs/2101.06804 arXiv:2101.06804 [cs].

[45] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *arXiv:2107.13586 [cs]* (July 2021). http://arxiv.org/abs/2107.13586 arXiv: 2107.13586.

[46] Vivian Liu and Lydia B. Chilton. 2021. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. arXiv:2109.06977 [cs.HC]

[47] Vivian Liu, Han Qiao, and Lydia Chilton. 2022. Opal: Multimodal Image Generation for News Illustration. 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 73, 17 pages.

https://doi.org/10.1145/3526113.3545621
[48] Vivian Liu, Jo Vermeulen, George Fitzmaurice, and Justin Matejka. 2022. 3DALL-E: Integrating Text-to-Image AI in 3D Design Workflows. arXiv:2210.11603 [cs.HC]
[49] Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. 2022. Things not Written in Text: Exploring Spatial Commonsense from Visual Signals. *ArXiv* abs/2203.08075 (2022).
[50] Maria Teresa Llano, Mark d'Inverno, Matthew John Yee-King, Jon Mccormack, Alon Ilsar, Alison Pease, and Simon Colton. 2022. Explainable Computational Creativity. *ArXiv* abs/2205.05682 (2022).
[51] Planetside Software LLC. 2022. Terragen. https://planetside.co.uk.
[52] World Machine Software LLC. 2022. World Machine. https://www.world-machine.com.
[53] Midjourney. 2023. Midjourney. https://www.midjourney.com/.
[54] Gaurav Mittal, Shubham Agrawal, Anuva Agarwal, Sushant Mehta, and Tanya Marwah. 2019. Interactive Image Generation Using Scene Graphs. *ArXiv* abs/1905.03743 (2019).
[55] Wizards of the Coast LLC. 2023. DnD Beyond. https://www.dndbeyond.com/.
[56] Siddharth Patki, Ethan Fahnestock, Thomas M. Howard, and Matthew R. Walter. 2019. Language-guided Semantic Mapping and Mobile Manipulation in Partially Observable Environments. In *Conference on Robot Learning*.
[57] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. 2019. Learn, Imagine and Create: Text-to-Image Generation from Prior Knowledge. In *Neural Information Processing Systems*.
[58] QuadSpinner. 2022. QuadSpinner Gaea. https://quadspinner.com.
[59] William L. Raffe, Fabio Zambetta, Xiaodong Li, and Kenneth O. Stanley. 2015. Integrated Approach to Personalized Procedural Map Generation Using Evolutionary Algorithms. *IEEE Transactions on Computational Intelligence and AI in Games* 7 (2015), 139–155.
[60] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *ArXiv* abs/2204.06125 (2022).
[61] Jeba Rezwana and Mary Lou Maher. 2022. Designing Creative AI Partners with COFI: A Framework for Modeling Interaction in Human-AI Co-Creative Systems. *ACM Transactions on Computer-Human Interaction* (2022).
[62] Bárbara Rodríguez-Fuentes and José Luis Ulloa. 2022. Why do people create imaginary worlds? The case of Fanfiction. *Behavioral and Brain Sciences* 45 (2022).
[63] 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]
[64] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *ArXiv* abs/2205.11487 (2022).
[65] Natalia Samutina. 2016. Fan fiction as world-building: transformative reception in crossover writing. *Continuum* 30 (2016), 433 – 450.
[66] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. 2016. Scribbler: Controlling Deep Image Synthesis with Sketch and Color. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 6836–6845.
[67] Shikhar Sharma, Dendi Suhubdy, Vincent Michalski, Samira Ebrahimi Kahou, and Yoshua Bengio. 2018. ChatPainter: Improving Text to Image Generation using Dialogue. *ArXiv* abs/1802.08216 (2018).
[68] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman. 2022. Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence. *ACM Trans. Comput.-Hum. Interact.* (feb 2022). https://doi.org/10.1145/3511599 Just Accepted.
[69] skeleton. 2023. *skeleton.* https://skeleton.dev
[70] SvelteKitC. 2023. *SvelteKit.* https://kit.svelte.dev
[71] Julian Togelius, Mike Preuss, Nicola Beume, Simon Wessing, Johan Hagelbäck, Georgios N. Yannakakis, and Corrado Grappiolo. 2013. Controllable procedural map generation via multiobjective evolution. *Genetic Programming and Evolvable Machines* 14 (2013), 245–277.
[72] Ruben Rodriguez Torrado, Ahmed Khalifa, Michael Cerny Green, Niels Justesen, Sebastian Risi, and Julian Togelius. 2019. Bootstrapping Conditional GANs for Video Game Level Generation. *2020 IEEE Conference on Games (CoG)* (2019), 41–48.
[73] Jia Wang, Owen Leach, and Robert W. Lindeman. 2013. DIY World Builder: An immersive level-editing system. *2013 IEEE Symposium on 3D User Interfaces (3DUI)* (2013), 195–196.
[74] Daijin Yang, Yanpeng Zhou, Zhiyuan Zhang, Toby Jia-Jun Li, and LC Ray. 2022. AI as an Active Writer: Interaction Strategies with Generated Text in Human-AI Collaborative Fiction Writing 56-65. In *IUI Workshops*.
[75] Zhenqiang Ying and Alan Conrad Bovik. 2020. 180-degree Outpainting from a Single Image. *ArXiv* abs/2001.04568 (2020).
[76] Jiahui Yu, Zhe L. Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. 2018. Generative Image Inpainting with Contextual Attention. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 5505–5514.
[77] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. *27th International Conference on Intelligent User Interfaces* (2022).
[78] Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. http://arxiv.org/abs/2106.11520 arXiv:2106.11520 [cs].
[79] Leah Zaidi. 2019. Worldbuilding in science fiction, foresight and design. *Journal of Futures Studies* 23, 4 (2019), 15–26.
[80] Jezia Zakraoui, Moutaz Saleh Mustafa Saleh, and Jihad Mohamad Jaam. 2019. Text-to-picture tools, systems, and approaches: a survey. *Multimedia Tools and Applications* (2019), 1–27.
[81] Chao Zhang, Cheng Yao, Jianhui Liu, Zili Zhou, Weilin Zhang, Lijuan Liu, Fangtian Ying, Yijun Zhao, and Guanyun Wang. 2022. StoryDrawer. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (2022).
[82] Lei Zhang and Steve Oney. 2020. FlowMatic: An Immersive Authoring Tool for Creating Interactive Scenes in Virtual Reality. *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (2020).

A APPENDIX

Topic	Design Prompt
Fantasy World	Imagine a world unlike any other, where the terrain is so varied and unique that each step you take leads you into a new adventure. The air is thick with mystery and magic, and the landscapes range from towering mountains to sprawling forests, vast deserts to shimmering oceans. Now, as a cartographer, it's your job to bring this world to life with your maps. With each stroke of your pen, you have the power to transport your readers to this magical realm and inspire them to explore every inch of its diverse and detailed landscapes. Are you ready to create a map that will take your readers on the adventure of a lifetime?
Landscape architect	As a landscape architect, your task is to create a fictional world that is a realistic simulation of the real world. The world should be designed with the same level of detail and accuracy as a real-life landscape, incorporating features such as realistic elevation changes, accurate water flow patterns, and vegetation that is appropriate to the climate and terrain
Computer Game	Create a computer game map that is both visually stunning and functionally complex, offering players a dynamic and engaging environment that requires strategic thinking and quick reflexes to outsmart opponents. Incorporate varied terrain, structures to capture and defend, hidden paths and secret locations, and balance gameplay mechanics to ensure a fair yet challenging experience. Transport players to a world of intense strategy and high stakes, challenging them to work together as a team to overcome their opponents in a game map that will be engaging, immersive, and memorable.
Treasure Hunting	Embark on an exciting adventure as a treasure hunter, exploring a map that is full of hidden riches and ancient artifacts waiting to be discovered. Use intricate puzzles, traps, and obstacles to create a challenging and engaging experience that will keep players on their toes. Populate the map with unique and exotic locations such as forgotten tombs, lost cities, and hidden underground chambers. Create a visually stunning landscape with diverse terrain, beautiful vistas, and atmospheric lighting that immerses players in a thrilling world of discovery and adventure.

Table A.1: An overview of the design prompts.

Code	Description	Examples
Size	Indicates relative sizes of the world's elements.	large; small; high; giant; tall; tiny; big
Positional	Terms that express how multiple objects are positioned in the image.	surround; above; side by side; around; underneath; middle; on both sides; bottom; left; corner; north; south; next to; in the carribbean; contain; in Rome; between; split by; inland; west; east
Action	Describes whether objects's actions.	hunting; selling; erupting; on fire; sits; running; coming; reach; wear; explosion; smoking; painting; holding; extending
Quantifier	Relates to the quantitiy of an object	many; few; dense; some; a lot of; lots of; singular; four; two; several
Style	Expresses the style of the image.	concept art; map; anime; cyberpunk; 1950; antique; cartoon; japanese; medieval; fantasy; futuristic; cartographic; geographical
Perspective	Relates to the viewpoint or perspective of the image.	2d; top down; horizontal; skyline view; view; isometric; bird

Table A.2: An overview of codes developed for the analysis of scene and region descriptions.

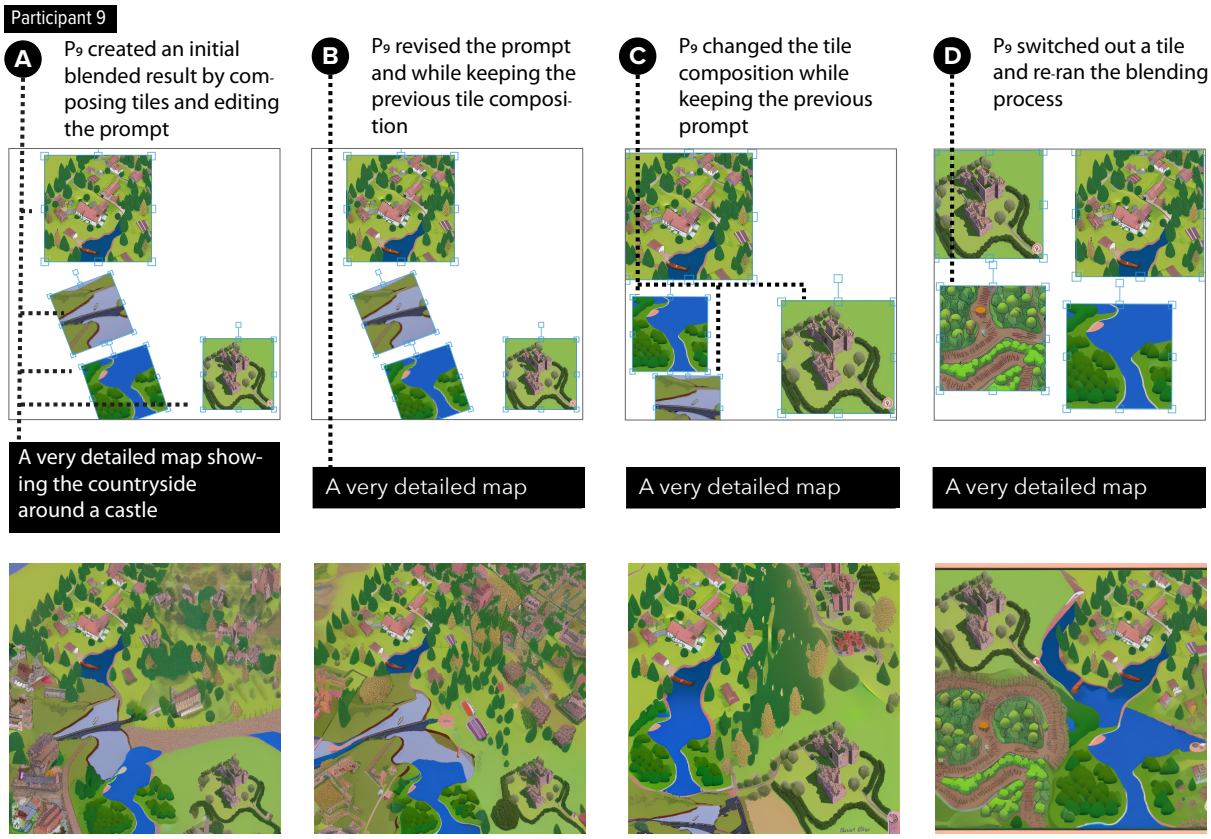

Figure A.1: Multiple world blending results with WorldSmith (P_9). (The top row) shows the four input tiles. (The bottom row) is the final blended image depicting participants fictional worlds. P_9 designed a top-down of a map consisting of a castle, a small village, farmlands, and a large river.