# Vistoria: A Multimodal System to Support Fictional Story Writing through Instrumental Image-Text Co-Editing

ANONYMOUS AUTHOR(S)

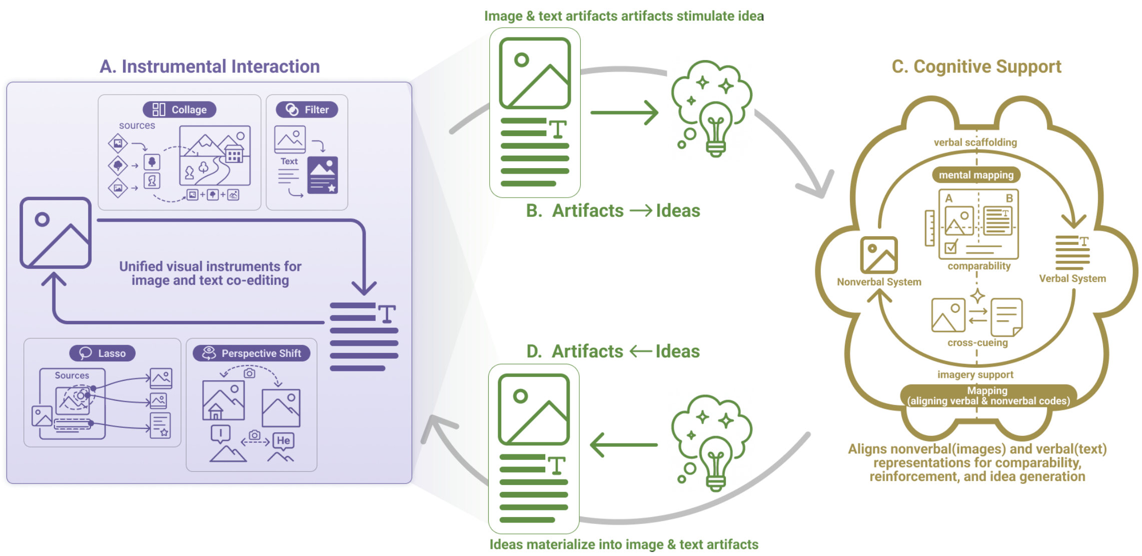



Fig. 1. Vistoria supports a cyclic workflow in which multimodal artifacts and ideas co-evolve. (A) Instrument Interaction: a unified set of instrumental operations (lasso, collage, perspective shift, and filter) enables image-text co-editing. (B) Artifacts ⟶ Idea: the resulting image-text alignment artifacts stimulate new story directions. (C) Cognitive Support: leveraging image–text alignment to synchronize verbal and non-verbal processing to enhance idea formation. (D) Ideas ⟶ Artifacts: emerging ideas are materialized back into new cards, closing the loop and driving iterative ideation, exploration, and integration.

Humans think visually—we remember in images, dream in pictures, and use visual metaphors to communicate. Yet, most creative writing tools remain text-centric, limiting how writers plan and translate ideas. We present Vistoria, a system for synchronized image-text co-editing in fictional story writing. A formative Wizard-of-Oz co-design study with 10 story writers revealed how sketches, images, and text serve as essential elements for ideation and organization. Drawing on theories of Instrumental Interaction, Vistoria introduces instrumental operations-lasso, collage, perspective shift, and filter that enable seamless narrative exploration across modalities. A controlled study with 12 participants shows that co-editing enhances expressiveness, immersion, and collaboration, opening space for writers to follow divergent story directions and craft more vivid, detailed narratives.. While multimodality increased cognitive demand, participants reported stronger senses of ownership and agency. These findings demonstrate how multimodal co-editing expands creative potential by balancing abstraction and concreteness in narrative development.

Additional Key Words and Phrases: Multimodality, Creativity Support, Storytelling, Creative Writing, Instrumental Interaction



Manuscript submitted to ACM

**ACM Reference Format:**
Anonymous Author(s). 2026. Vistoria: A Multimodal System to Support Fictional Story Writing through Instrumental Image-Text Co-Editing. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Creant el demà junts, Barcelona.* ACM, New York, NY, USA, 35 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 Introduction

Human thinking involves multimodal processing. Visual processes play a central role in cognition: we recall experiences as spatial scenes, form mental models through imagery, and use visual structure to organize and interpret information [36, 37, 68]. Language is similarly entwined with imagery. Particularly, text comprehension often evokes mental pictures, and abstract ideas are commonly articulated through spatial metaphors such as path, framework, or perspective [55]. Dual Coding Theory frames this coupling between imagery and language, positing that humans draw on both verbal and nonverbal channels to represent and support specific reasoning and communicative processes [20].

Story writing is particularly multimodal in nature. During the planning phase, experienced writers often use both imagery and language to construct the story world. They visualize spatial layouts, character interactions, and scene dynamics, while using textual notes to label, sequence, and reason about narrative structure [4, 19, 22, 25]. In the translating phase, visual details serve as an anchor that shapes how writers organize narrative detail and emotional tone, while texts linearize these visualized ideas into descriptions, dialogue, and narrative perspective that readers can follow [49, 61]. Crucially, nonverbal and verbal channels do not operate in isolation: writers use imagery to trigger new wording, and emerging text in turn elicits further mental images [43, 88]. Therefore, there is a great opportunity for tools that support story writing to match this multimodal complexity by accommodating the continuous interplay between visual and textual thinking within the same workflow.

Yet, current writing tools remain overwhelmingly text-centric, treating linear text as the primary or sole medium of expression [31, 44]. Although recent systems powered by Large Language Models (LLMs) incorporate visual elements through image generation or retrieval, these visuals remain peripheral. They function mainly as prompts [66], static references [67], or organizational diagrams [59, 80], rather than as tightly integrated, manipulable representations along with text. This requires writers to translate visual ideas back into text, increasing cognitive load [11, 85, 88]. Such back-and-forth conversion also constrains cross-modal collaboration and narrows the range of creative possibilities that a unified multimodal system could otherwise support.

To examine this gap, we conducted a formative Wizard-of-Oz (WoZ) co-design study with 10 experienced writers to inform the design of a unified image-text multimodal system that supports the planning and translation phases of fictional story writing. We found that text and images play distinct yet complementary roles in the writing process. Writers expressed a need to directly manipulate text and images for fine-grained editing and alignment, valuing the ability to move fluidly between them.

Based on these formative results, we developed Vistoria, a system that transforms fictional story writing from a text-centered process into a multimodal co-editing experience that integrates text, images, and sketches. The design of Vistoria draws on the principle of Instrumental Interaction, designing a set of instrumental operations (lasso, collage, perspective shift, and filter). These functions can be applied to either text or images, where a single action simultaneously affects both images and text, minimizing switching costs and preserving creative flow [6, 69, 72]. Based on Dual Coding Theory, Vistoria enables the alignment of text and visual representations, ensuring that edits in one modality are appropriately reflected in the other.

We conducted a controlled study with 12 participants to examine how multimodal image–text co-editing supports fictional story writing, with a focus on evaluating the system's usability and understanding its creative support. Overall, the study showed that Vistoria enhances expressiveness, immersion, and exploration, enabling participants to have more divergent ideas and write detailed narratives. Participants used the instrumental operations to refine ideas at multiple scales, explore alternative directions. While this workflow increased mental and physical workload, it also supported writers' senses of agency and ownership, as they maintained greater operational control over narrative development.

In summary, this work contributes:

- A WoZ co-design study with 10 writers examined the practices and needs of using multimodal elements to externalize ideas and develop narratives in the planning and translating phase of fictional story writing;
- Vistoria, a multimodal co-editing system that unifies text and visual images through instrumental operations and a synchronized editing loop to support fictional story writing;
- A controlled usability study with 12 participants demonstrates the potential of Vistoria, suggesting that multimodal co-editing can enhance expressiveness, idea generation, and narrative development in fictional story writing.

## 2 Related Work

### 2.1 Using Visuals to Support the Cognitive Process of Fictional Story Writing

Fictional story writing is distinct from argumentative or expository genres in its emphasis on imagination, world-building, and character development [25]. Writers must invent narrative worlds and characters while ensuring coherence, which poses unique cognitive challenges: abstract, nonverbal mental images must be transformed into structured narrative elements and then into text [4].

The Cognitive Process Model of Writing [28, 30] frames writing as recursive processes of planning, translating, and reviewing. In fictional story writing, the phases from planning to translation are especially demanding, as writers move from imaginative constructs to linear verbal representation, imposing a high cognitive load due to the need for simultaneous translation and structural organization. However, visual representations can scaffold this process by externalizing abstract ideas. Research shows that picture prompts improve writing coherence [61], and visual images stimulate creativity in narrative writing [49]. In practice, sketches, maps, and diagrams externalize plot, setting, and character relationships that writers actively manipulate during planning and revision, while also serving as cognitive anchors during translation to maintain coherence and consistency. [88]. Dual Coding Theory [20] explains these benefits: verbal and nonverbal systems function separately but also interact, creating richer memory traces when information is encoded in both modalities. In fictional writing, visual representations of narrative elements complement verbal planning, making abstract concepts more concrete and retrievable. When writers encounter difficulties in translation, visual anchors provide alternative access to imaginative content, reducing cognitive load and enabling more fluid expression [5]. These visual structures are not merely supportive; they function as alternative representational spaces in which writers perform cognitive operations that parallel textual editing and support non-linear narrative leaps [73, 85].

However, existing creativity-support systems largely leverage visuals in limited ways, focusing on inspiration, reference, or structural overview rather than enabling writers to directly manipulate visual narrative elements and align with text editing. Planning-focused systems such as CCI, Sketchar, and CharacterMeet assist authors in character and world development, through image-guided backgrounds or conversational refinement of characters [48, 65, 66].

Translation-focused tools like ScriptViz and Script2Screen aim to align textual composition with visual referents, either by retrieving reference visuals from movie databases [67] or by synchronizing scriptwriting with audiovisual scene creation [82]. Complexity management systems, for example, WhatIF [59], ClueCart [80], and PlotMap [81], help writers maintain structural coherence by visualizing branched narratives, organizing narrative clues hierarchically, or integrating spatial layouts with textual plot structures.

In these systems, writers may look at images to spark ideas, but operations such as cutting, re-ordering, or reframing narratives still have to be performed only in the verbal channel. Because the underlying story state is effectively defined only through text, visuals cannot serve as core writing operations such as restructuring events, adjusting focalization, or reorganizing character relationships [28, 88]. This separation requires writers to repeatedly convert visually grounded ideas back into verbal form for any narrative change to take effect, thereby increasing cognitive load and undermining many of the well-established benefits of external representations such as diagrams, sketches, and other forms of external cognition [40, 74]. As a result, images remain outside the recursive planning–translating loop. To address this gap, our work treats visual and text representations as synchronized and co-editing materials, allowing writers to manipulate narrative elements across image and text through the same set of operations and thereby more tightly aligning verbal and nonverbal with the cognitive processes of fictional story writing.

### 2.2 LLM-powered Multimodality Tools for Creativity in Content Creation

Recent multimodal creativity tools move beyond linear prompting by enabling direct manipulation of creative elements, helping creators express intentions that language alone cannot capture [52, 69]. Powered by advanced Large Language Models (LLMs), these systems address fundamental barriers through three complementary mechanisms. First, they externalize creative structures, and support better intention expression. Tools like AI-Instruments [69] and Brickify [72] transform abstract intentions into manipulable interface objects or reusable visual tokens, rendering otherwise ineffable ideas as visible, persistent, and operable elements. Second, in recent multimodal systems, sketches always act as a nonverbal, spatially grounded modality that conveys structure, hierarchy, and relations far more efficiently than language. DrawTalking [70] combines freehand sketching with spoken narration, enabling natural intention communication. Code Shaping [86] allows developers to make sketched annotations directly on the code editor to support fuzzy, incremental expression of intent. Third, supporting iterative refinement, Inkspire [47] and AIdeation [79] accelerate variation and exploration, enabling rapid cycles of sketch-to-output or recombination of references.

Seeking tighter coupling for fictional story writing, recent systems push the integration of multimodal interaction in different ways [18, 19]. These systems are designed in response to growing evidence that text-only, model-driven workflows cause LLM-assisted stories to converge toward similar narrative structures, limiting exploration, reducing originality, and diminishing writers' expressive control [13, 24, 46]. WorldSmith supports layered edits and hierarchical compositions through sketches, making it easier to grow a world piece by piece instead of through single and monolithic prompts [22]. XCreation [85] supports cross-modal storybook creation by integrating an interpretable entity-relation graph, improving the usability of the underlying generative structures. Toyteller [19] maps symbolic motions to character actions, letting users express rich social and emotional interactions that are often hard to write down explicitly only using text. Visual Writing defines an approach where writers edit stories by manipulating visual representations to make the underlying narrative structure more comprehensible and easier to work with than linear text alone [53].

This line of multimodal research demonstrates that combining language with gestures, sketches, and direct manipulation can offload cognitive work from linear prompting and give creators more expressive, situated channels for specifying and revising intent. Building on this line of work, our system introduces a canvas-based interface that integrates images, sketches, and text to support the externalization of mental imagery and the articulation of narrative intent. Through designing a set of instrumental operations for image–text co-editing to enable iterative refinement, Vistoria allows writers to move fluidly between verbal and nonverbal modes of thinking as they develop fictional stories.

### 2.3 Instrumental Interaction

Instrumental Interaction is central to understanding how users control and refine digital systems. Beaudouin-Lafon [6] proposed it as a shift from designing static interface elements to designing instruments that mediate between users and domain objects. A key principle is reification, which transforms abstract commands into persistent, manipulable objects [39]. In computing, this elevates implicit system descriptions into explicit first-class entities. In LLM-assisted workflows, this principle appears in modular prompt blocks for structured edits [87] and in Textoshop's reification of abstract image editing commands (e.g., tone adjustment, boolean operations, layers) into direct manipulation tools for text [52]. A second principle is polymorphism, where the same instrument applies across contexts [51]. This reduces cognitive load by enabling predictable, transferable patterns, e.g., copy–paste works consistently across text, images, and files [1], and scrollbars operate similarly across documents, spreadsheets, and browsers [51]. Finally, reuse allows users to replay or adapt prior operations, from macros to redo commands [69]. Systems like Spacetime exemplify this by objectifying space, time, and actions into persistent containers, enabling edits to be carried forward as manipulable entities [83]. Together, these principles reduce cognitive burden by externalizing interaction histories, making them manipulable, transferable, and extensible.

In our system, we extend this perspective to the design of multimodal tools for fictional story writing. We reify narrative development as a set of instrumental operations spanning text and imagery. Through polymorphism, the same operation can be applied to both LLM-generated text and images. This combination of reification and polymorphism enables writers to shape multimodal outputs fluidly, aligning with both verbal and nonverbal perceptions.

## 3 Formative Study

Previous research shows that multimodal tools enhance fictional story writing by making abstract concepts tangible, reducing cognitive load, and improving creativity and coherence [17, 18, 22, 88]. However, current tools treat images as supplementary rather than integral to the creative process, leaving unclear how writers actually integrate multiple content types into a cohesive workflow.

To address this gap, we conducted a WoZ co-design study [21] examining how creators use multimodal content (images, text, sketches) when planning and drafting fictional stories [28, 30, 88]. Our investigation focused on three questions: (1) **Multimodal information use:** what types of multimodal content users employ and how they leverage these materials for idea generation; (2) **Iteration and integration:** how creators refine and combine multimodal artifacts in world-building and narrative development; and (3) **Organization of inspirations:** how creators organize, connect, and refine dispersed inspirations through multimodal manipulation. The WoZ setup simulated AI-assisted visual and textual support while sustaining the impression of an intelligent, interactive system.

Our system is designed for writers with intermediate to expert writing expertise, rather than novices who are still learning basic narrative composition. This target group typically possesses established writing habits and a solid understanding of narrative structure. As contemporary writers increasingly incorporate large language models (LLMs) into their creative workflows (idea generation, style adjustment, etc.) [14, 43], we target writers who have hands-on experience using LLMs to assist their writing, even though they may not be experts in multimodal interaction or prompting.

## 3.1 Process

We designed a Wizard-of-Oz (WoZ) co-design study, positioning participants as active co-designers and treating text, sketches, and images as shared design materials [21, 71, 78].

*3.1.1 Participants.* For the formative study, we recruited 10 participants through student organizations by sharing our study announcement in group chats, each with at least two years of experience in creative writing. The group included three fictional story writers, three animation scriptwriters, two visual film creators, one new media creator, and one online fiction writer. Eight participants held a master's degree or higher, and two held a bachelor's degree. All participants had experience using LLMs to assist in their writing, AI familiarity ranging from casual use (fewer than two days per week, n=5) to daily workflow integration (five or more days per week, n=5).

*Experimental Setup.* Three days before the session, participants were instructed to prepare a brief fictional story outline consisting of several sentences that followed one of the narrative structures from The Seven Basic Plots [9], which served as the foundation for subsequent ideation and content development. The 90-minute main session took place in either Figma [27] or Miro board [58], based on participant preference. Each session concluded with a 30-minute semi-structured interview probing how multimodal materials mediated co-creation, and what interaction patterns and workflows participants desired.

*3.1.2 Wizard-of-Oz System and Session Process.* For the WoZ interface, we utilized the canvas in either Figma [27] or Miro [58] as a collaborative space, which was divided into (1) a user-facing "Text Editor" where participants can put in the outline and edit the story, (2) the "Canvas" where generated images and text, and participants' notes were, and (3) a hidden "Wizard Control Center" (as shown in Figure 2). Participants communicated via voice, text (stick notes), or hand-drawn sketches while two researchers acted as the "Wizards (system backend)" in real-time to generate outputs, ensuring responsiveness, copying user inputs into separate windows. Researchers ran Claude for text generation, ChatGPT (GPT-4o) and Midjourney[1], for simulating the visual engine, then pasting the results back onto the "Canvas". To ensure consistency, the Wizards followed: (1) input the user's sketch/text as a literal prompt. Use the initial outline as contextual information for prompts; (2) do not offer creative suggestions unless explicitly asked; (3) to ensure diversity of output style, one researcher generated both texts and images via ChatGPT, and the other researcher generated images via Midjourney and texts via Claude. The two wizards ensured that participants were provided with results that were both timely and diverse.

*Session Process.* During the session, participants engaged in fictional story co-design through activities including generative prompts, collage, and storyboard-like arrangement while interacting with the wizards. Rather than working toward a fixed output, participants iteratively developed stories of about 300 words while envisioning how the tool itself should behave.

## 3.2 Formative Study Findings

*3.2.1 Using Multimodal Input to Reify Vague Ideas.* As shown in the **Appendix 2**, participants utilized multimodal expressions, including text, sketches, and images, as co-design materials to articulate and negotiate intentions with the wizarded system.

**Sketches** externalized vague intentions and spatial imagination. For example, P4 envisioned a scene where clown Joko appeared on stage and created a sketch with textual annotations describing the intended atmosphere, hoping AI

[1]All models were accessed via their commercial web interfaces: https://claude.ai [2], https://chatgpt.com [62], and https://www.midjourney.com/ [57] respectively, in June 2025.

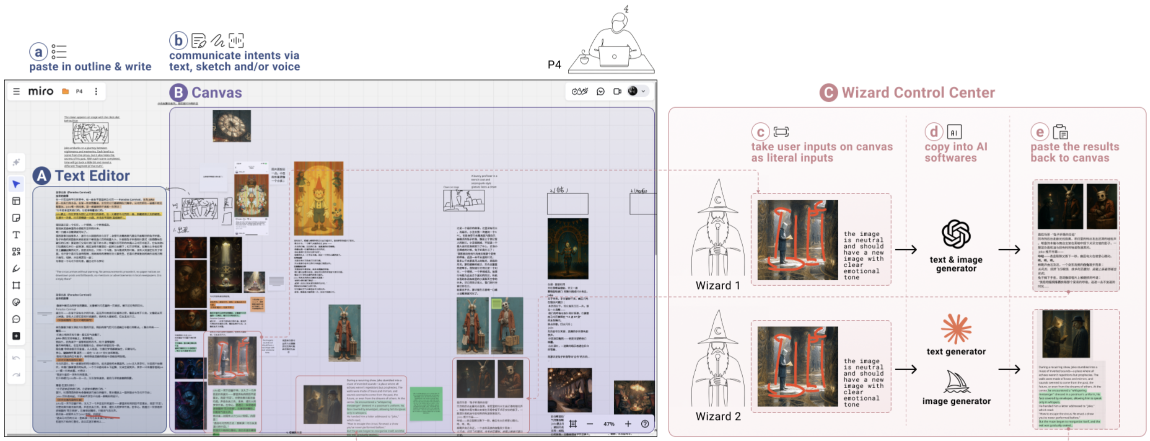



Fig. 2. WoZ System and Study Process (the example of P4) (a) The user edits text in the Text editor; (b) The user writes, sketches, and speaks out their intents; (c)-(d) both wizards paste in the user's inputs into AI software windows; (e) wizards pasted the generated results back onto canvas.

could elaborate narrative details. Sketches were also frequently used to express ideas that participants found difficult to convey through prompts alone (P6, P7, P8), as they were seen as carrying richer layout and spatial information. **Text** functioned as the primary medium for conveying intent, allowing participants to express connections between desired content and existing stories. Participants also frequently used textual annotations to specify story parts or image types for AI-generated content and to guide the direction of content generation. They also used textual annotations to record the inspirations they received and how those ideas might be used in writing. **Images** communicated style and mood expectations. P3 altered an AI-generated picture's style by supplying a reference image, while P1 noted, *"if possible, I want to use a 'supporting image'—a vague reference picture—as a basis, expecting the AI to generate more detailed images derived from it."* Combining image outputs with textual descriptions helped participants enrich their limited knowledge. For example, P7 requested AI-generated designs of an ancient Chinese poison bottle as a narrative element, noting her vague understanding of the concept. In addition, sketches were often paired with images or text to further articulate the intentions participants held in mind (P4, P7).

For the content, participants most often sought LLM elaboration on **characters**, **objects**, and **scenes**, expressing the need for assistance with character design refinement, setting depictions, or object visualization. These findings highlighted the value of systems that accept multimodal input and help co-designers transform nascent ideas into concrete narrative materials.

*3.2.2 Image-Text Interplay as Complementary Design Moves.* Participants perceived text and images as distinct yet complementary in their story writing. We observed participants often switching between abstraction (text) and concreteness (images) as a recurring co-design pattern.

*Text as open imagination.* Participants described text as a "blank canvas" for boundless imagination. P4 noted, *"text allows me to imagine many things in my mind,"* and P1 emphasized text as "infinite imagination on a blank page." P8 highlighted that text helped set up the narrative structure before layering in visuals. This suggests systems should treat text as a flexible space for ideation and intent communication, where ambiguity can be preserved rather than early resolved.

*Images as concreteness, inspiration, and feedback.* Images grounded abstract ideas, while their randomness often sparked unexpected inspiration. P1 explained: *"The randomness in AI-generated images goes beyond what I want or can express; it helps me imagine the next step of the story."* Similarly, P4 refined Lily's behavior based on an unexpected visual detail, and P3 used images as feedback for progressive refinement. P10 regarded referring to images as a "look-and-write exercise" that scaffolds scene construction. These accounts highlight the value of image outputs not just as illustrations but as provocations. participants can leverage it through image-based iteration, selection, and reinterpretation.

*Complementary interplay.* Participants emphasized that neither modality sufficed alone: images "set the vibe," while text reframed meaning. P8 noted, *"Images can serve as references for appearance when I don't have many ideas, while text quickly triggers associations."* We observed participants iteratively moving between text for open-ended imagination and images for concrete grounding, forming a cycle of divergence and convergence (P4, P5, P8, P9, P10). Four participants (P1, P2, P5, P7) also expressed a desire for text and image changes to be synchronized, so they would not need to constantly cross-check and compare updates across modalities. In addition, the concurrent presentation of text and images further facilitates narrative expression. As P7 explained, *" having text and images appear together allows inspiration to arise simultaneously in both modalities, helping me understand how to describe the scene more effectively. I can describe the scene while referencing the image, and directly adopt AI-generated text when I find it useful."* They noted that when an image is edited, the corresponding text should update accordingly.

This interplay suggested design opportunities for systems that incorporated smooth transitions between text and images while aligning these two modalities, enabling participants to fluidly move between abstract exploration and concrete elaboration

*3.2.3 Direct Manipulation of Multimodal Artifacts.* Participants expressed a strong interest in treating text and images as manipulable, recombinable design materials. Two recurring practices pointed to design needs for more fluid multimodal manipulation.

*Collaging and Recombination.* Participants frequently merged elements across outputs to spark new ideas. P9 envisioned combining "the house from the first AI-generated images with the street from the second picture" to construct scenes, while P7 highlighted that "randomly combining characters and scenes" could inspire unexpected connections when accompanied by textual descriptions. As she explained, *"if I can see an AI-generated image of my protagonist in one of the scenes, it helps me better imagine potential connections between elements that might otherwise seem unrelated. Being able to visualize scenes that are otherwise difficult to imagine enables me to write more narrative descriptions more easily."* Such practices illustrated the potential of collage and recombination as creative strategies.

*Granular Editing and Annotation.* Beyond recombination, participants desired fine-grained control over outputs. Sticky notes captured details for iteration and served as prompts for later development. Participants also wanted more localized operations, such as regenerating specific regions (P1, P3), extracting and reusing circled image elements (P5), or annotating character personas for refinement (P2). They left narrative prompts for later translation, e.g., P2's note *"Ending could be related to why this postman job even exists."* These behaviors emphasized editing and annotation as both vehicles for iteration and a bridge to subsequent writing.

Together, these findings suggested systems should enable flexible recombination, localized editing, and traceable annotations to help creators iteratively refine narrative materials.

*3.2.4 From Fragmented Inspirations to Coherent Storylines.* While participants often highlighted text or circled inspiring image details, organizing these dispersed fragments into coherent narratives was a persistent challenge in the Planning phase. As P2 noted, *"everything quickly became too messy on the canvas,"* and P4 likened fragments on the canvas to "many cards that required connections," where narrative coherence depends on linking passages, characters, and settings from scattered parts. Furthermore, P6 wished for mind map–like tools to scaffold this process. Participants also requested clustering notes, surfacing latent relations (e.g., by character/object/setting), and consolidating materials into reusable "setting cards" to ensure cross-chapter consistency and avoid logic conflicts (P4, P7).

These challenges pointed to opportunities for systems that transform fragmented inspirations into structured storylines by supporting clustering, relation mapping, and the creation of reusable narrative units that preserve coherence across iterations.

### 3.3 Design Goals

Drawing on insights from our WoZ co-design study, prior work on multimodal LLM tools, and theories of Structural Mapping and Instrumental Interaction (Section 2.3), we identify four design goals for a multimodal content creation interface that supports the planning and translating phase in fictional story writing [28, 30, 88].

- **DG1: Supporting the Expression of Ideas through Multimodality.** Grounded in the findings in Section 3.2.1, our system should provide multimodal mechanisms combining sketches, text, and images to capture early intention, imprecise expressions, and help transform them into concrete narrative materials for further iterative editing or re-organizing.
- **DG2: Aligning Text and Images for Iterative Creative Exploration.** Informed by findings in Section 3.2.2, our system should enable fluid cross-modal iteration: textual edits can be re-visualized, and image refinements can inform text descriptions. Grounded in Dual Coding Theory, text and image updates should also be synchronized to more effectively align verbal and nonverbal perception.
- **DG3: Enabling Polymorphic Cross-Modal Manipulation.** Informed by findings in Section 3.2.3, our system should support direct manipulation interactions for both text and images. Guided by Instrumental Interaction's principle of polymorphism, we should design the same instrument for cross-modal editing to reduce switching costs and enable writers to manipulate textual and visual fragments while maintaining narrative coherence.
- **DG4: Organizing and Reusing Fragments into Coherent Narratives.** Informed by the findings in Section 3.2.4, our system should support clustering and organizing fragments and fleeting ideations during the exploratory phase, surface latent connections, and consolidate dispersed inspirations into coherent, evolving narrative structures that support translation into final writing.

## 4 Vistoria System

In this section, we present the key features of Vistoria. As shown in Figure 3, the interface comprises three primary components: a left text editor, a central collapsible cluster panel, and a right canvas interface. The text editor displays the current story draft, serving as contextual information for content generation. The right canvas supports freeform sketching, text input, and image-text generation and editing tools. The central cluster panel aggregates highlights and annotations from canvas, displaying related plots, settings, and descriptions of each highlighted element for easy reference and overview.

### 4.1 Key Features

*4.1.1 Reifying Intention through Multimodal Generation.* The system enables writers to externalize early, vague ideas using multimodal inputs (DG1). To support this, Vistoria converts multimodal inputs into cards that pair an image

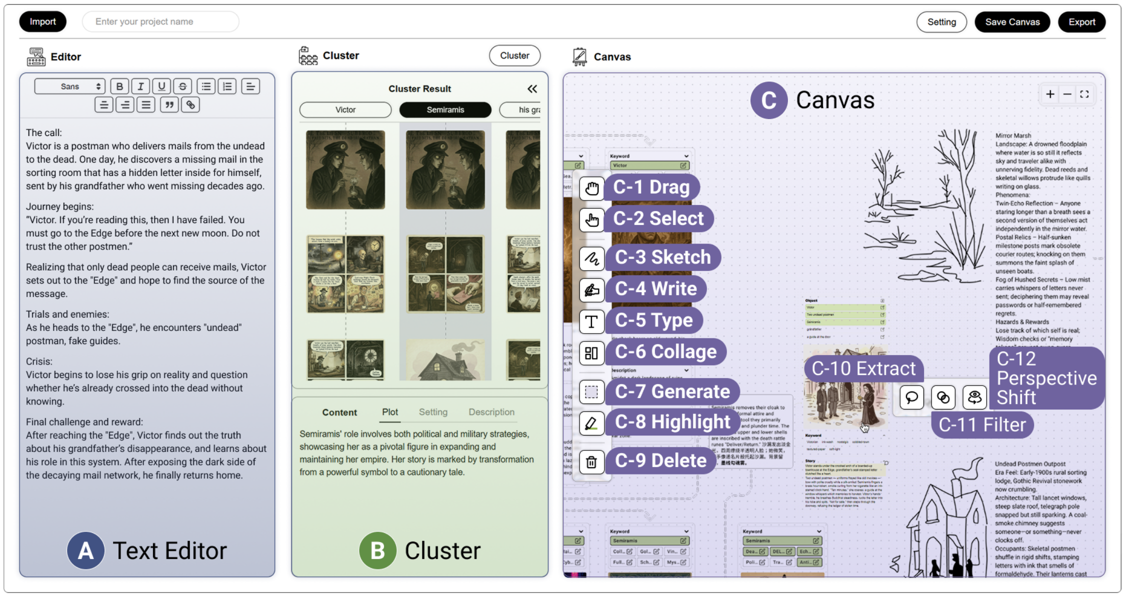



Fig. 3. Vistoria's interface: a left text editor, a central Cluster panel (can be collapsed when not used), and a right free-form Canvas.

with a narrative segment. These cards reify multimodal input into reusable artifacts that persist on the canvas and can be iteratively regenerated, while also serving as alignable units in which text and image convey the same underlying meaning.

Vistoria further balances precision and exploration by offering two complementary generation modes. In *Exact Craft* mode, single cards closely adhere to the author's expressed intention to concretize specific ideas. In *Creative Spark* mode, three cards are generated to represent diverse options based on the writer's intention. The system deliberately introduces variation around characters, settings, or objects, providing alternative prompts that can inspire new directions.

*4.1.2 Synchronized Image–Text Co-Editing through Instrumental Operations.* Fictional story writing benefits from fluid movement between abstract textual reasoning and concrete visual imagination. The system should tightly align text and images so that edits in one modality fluidly inform the other (DG2, DG3). To address this, we introduce a set of *Instrumental Operations* (Figure 4) designed around three principles: (1) *Reification (instrumental interaction)*, which draws on familiar image-editing operations to make abstract image–text co-editing actions more concrete and manipulable; (2) *Polymorphism (instrumental interaction)*, designing a set of instrumental operations (lasso, collage, filter, perspective shift) which ensure the same operations apply uniformly across text and images to lower switching cost; and (3) *Dual Coding Theory*, which indicates that verbal and nonverbal changes should be aligned to maintain coherence across cognitive channels.

***Lasso.*** The *lasso* instrument exemplifies reification by turning the abstract action of "focusing on part of a story" into a manipulable unit: selecting a region in *either* an image or a fragment of text triggers the generation of a new card focusing on the selected part with enriched narrative and visual details. Through polymorphism, the same selection logic applies across modalities—whether circling a visual detail or isolating a text segment—providing a consistent

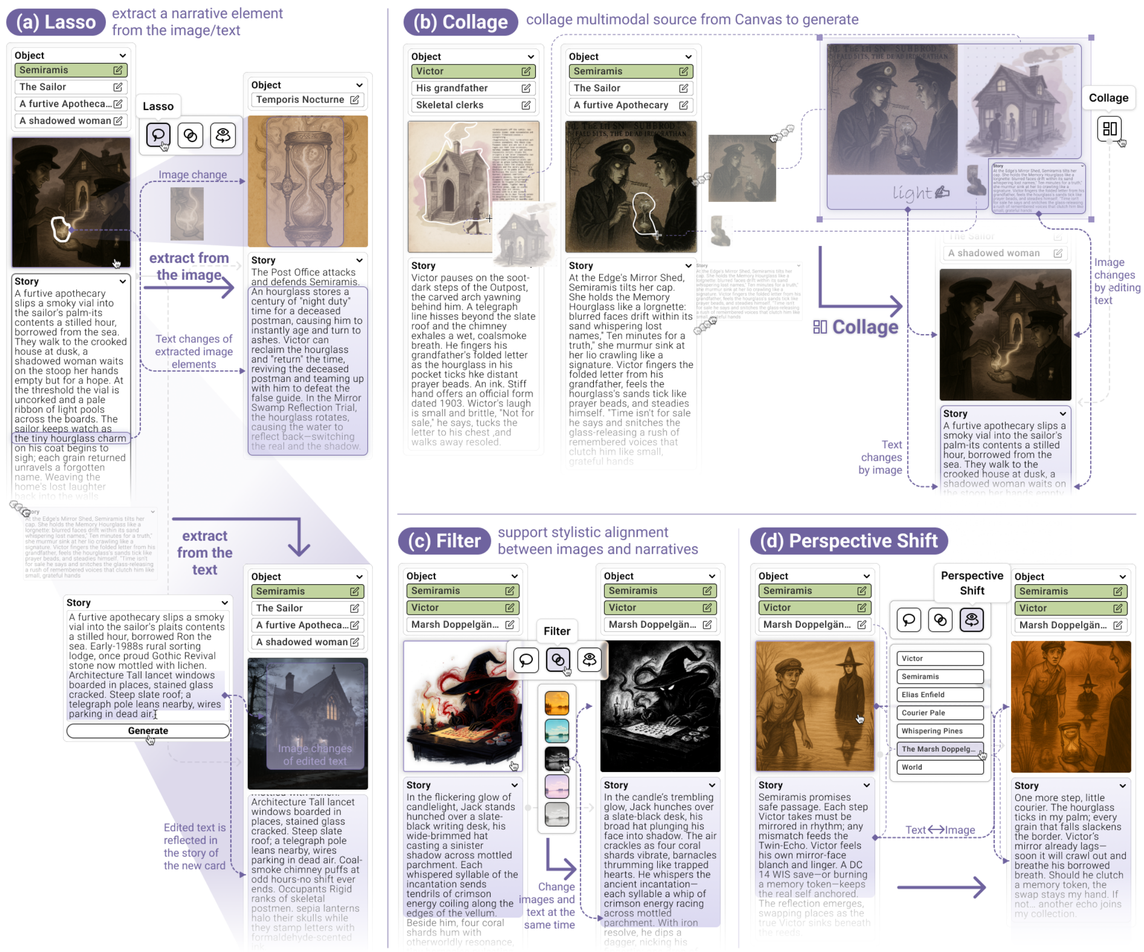



Fig. 4. A set of instrumental operations for image-text co-editing to enhance planning and translating of fictional story writing: (a) Lasso selects regions for coupled image–text edits. (b) Collage enables writers to extract elements and compose across cards to discover new narrative directions. (c) Perspective Shift changes an image's viewpoint and automatically regenerates the story's point of view (first/third/second person). (d) Filters align visual style and textual tone (e.g., melancholic/dreamy) by jointly altering image effects and rewriting prose.

interaction pattern. The text within the selected area in the original content will be emphasized to form a new card. The extracted portion of text is used to regenerate the corresponding image. The lassoed image region and the expanded story fragment correspond to one another, aligning visual and textual perspectives within the same narrative unit (Figure 4 (a) ).

***Collage.*** The *collage* instrument reifies the abstract act of "recombining inspirations" into a tangible manipulation: fragments of images, sketches, or text can be directly composed within a collage frame to form a new card. The same cut–paste–combine logic applies uniformly across modalities—an image region, a text excerpt, or a sketch element can all be treated as compositional materials for intention-based generation. The system interprets the spatial arrangement of these multimodal pieces as narrative intent, generating a card where textual descriptions and visual depictions are aligned. For instance, merging two character fragments not only produces a combined image but also generates a new story segment situating them together, ensuring that narrative and imagery evolve in sync (Figure 4 (b)).

***Filter.*** Stylistic coherence is critical in fictional story writing, as consistent affective and aesthetic cues sustain narrative transportation [32], activate readers' interpretive schemas [3], and enhance the emotional resonance of literariness [56]. In Vistoria, the *filter* instrument reifies this abstraction into a concrete tool: applying a "melancholic" or "dreamy" filter adjusts the visual style and rewrites the accompanying prose to match the emotional tone (Appendix Table 4). Through polymorphism, the same filter operation works seamlessly across modalities, leveraging the correspondence between visual style in images and emotional tone in text to simultaneously act on both. By making intangible stylistic intentions manipulable and synchronized, filters expand expressive possibilities while maintaining narrative immersion (Figure 4 (c)).

***Perspective Shift.*** Fictional story writing often utilizes perspective shift, and narratology highlights that changes in voice and focalization fundamentally reshape how events and characters are perceived [29]. Cognitive poetics further shows that such shifts alter readers' empathy and immersion. First-person narrations foster intimacy, while third-person perspectives enable broader structural awareness [38]. The perspective-shift instrument reifies this narratological concept into an actionable operation: changing the visual viewpoint of a scene automatically regenerates the story fragment from a first-, third-, or second-person perspective. Through *polymorphism*, this instrument applies consistently across modalities, altering either an image or its accompanying text triggers a corresponding adjustment in the other. The shift carries the same meaning across text and image: a new camera angle in the image corresponds to a new narrative voice in the text, allowing writers to explore empathy, distance, and awareness in a synchronized manner (Figure 4 (d)).

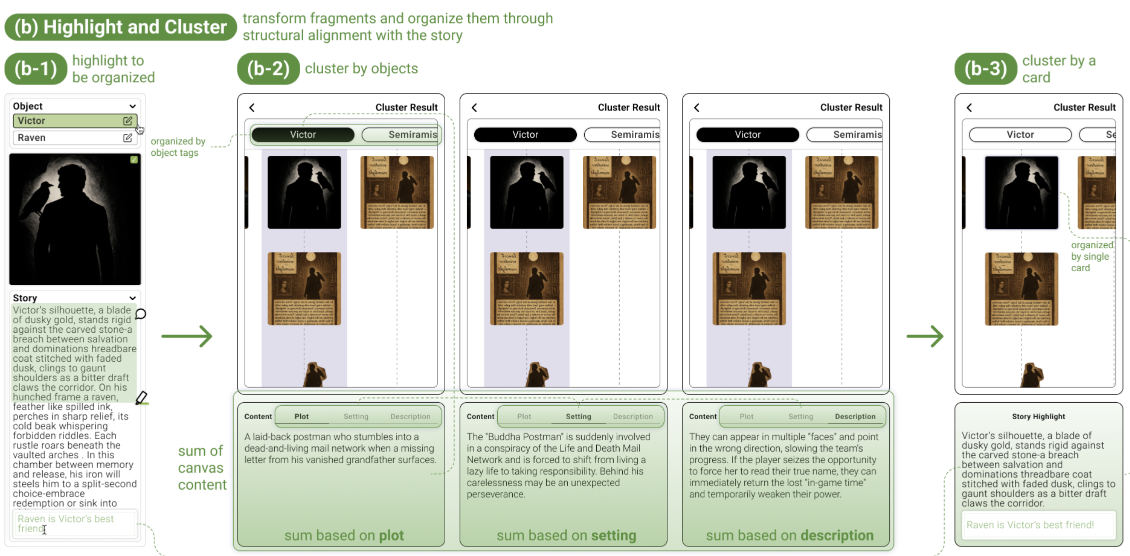



Fig. 5. Writers highlight objects and text segments on cards (b-1); the Cluster panel aggregates these by character/object/scene and can auto-summarize settings/plot/description about a certain object to guide final writing (b-2); Clicking on a specific image reveals the corresponding highlights and comments from earlier phases left on canvas (b-3).

*4.1.3 Highlight Elements and Cluster.* Writers often struggle to integrate scattered highlights and annotations on the Canvas into coherent storylines, leaving ideas fragmented across cards (DG4). Vistoria addresses this by transforming the dispersed fragments into reusable narrative building blocks, aligning them structurally across characters, objects, and scenes. This process is centered in the cluster panel (Figure 5), which turns fragmented inputs into organized knowledge assets.

On the canvas, writers can highlight textual segments, edit stories directly, and add inline comments noting potential uses in later drafting. Story objects, such as characters, settings, or scenes, are represented as editable keywords that can be highlighted by themselves. The system automatically links each highlighted object to its associated text, consolidating references across multiple cards.

The cluster panel then aggregates all highlighted objects into an organized overview of evolving narrative elements. This eliminates the need for manual scanning of scattered cards and provides writers with a dynamically updated, object-centered workspace. Selecting an object reveals its complete set of associated materials, including linked images, highlighted text segments, and comments, which creates a multimodal, context-rich reference for downstream writing. Beyond simple aggregation, the panel supports higher-level knowledge construction through its summary feature. When this feature is invoked, the system generates structured summaries of settings, descriptions, and plot elements derived from highlights and comments. These summaries distill fragmented annotations into narrative building blocks [12], enabling writers to iteratively scaffold coherent storylines from previously disjointed ideas.

## 4.2 Implementation

We adopted a decoupled front-end/back-end architecture. The React[2] front-end enables efficient rendering for complex interactive interfaces, while the Flask[3] back-end flexibly handles model calls with minimal overhead. Axios manages asynchronous communication between layers.

*4.2.1 Front-end.* The front-end consists of three main modules: the Canvas, the Cluster, and the Text Editor. Zustand[4] centrally manages the global state (including canvas nodes, cluster selections, and text content) to ensure consistency across all modules. To protect privacy during user studies, per-session data is stored in sessionStorage and automatically clears when the tab closes, while users can manually export canvas nodes and text content via the top toolbar.

*Canvas Module.* The right-side canvas module consists of four distinct layers: (1). Node Interaction Layer (Bottom): This layer uses React-Flow[5] to maintain a dynamic node-edge graph. Node types include card, collage, text, sketch, handwriting, and image. All nodes share basic properties, such as ID, type, and coordinates, for edge linking, but the internal data structures of those node types vary for rendering. For example, the "card" node includes features such as image modification tools, image lasso selection, image highlighting, object manipulation, and basic information display. (2). Pen-based Input Capture Layer: This layer supports natural interactions like freehand sketching, writing, and lasso selection. It uses the perfect-freehand library[6] to smooth captured points and convert them into Scalable Vector Graphics (SVGs). Each captured SVG stroke is added to the graph as a new node. (3). Generation Selection Layer: A Document Object Model (DOM) based screenshot function ensures visual and positional consistency. After clicking generate, the front-end sends both the screenshot and structured data of the nodes inside the selected area to

[2]https://react.dev/
[3]https://flask.palletsprojects.com/
[4]https://github.com/pmndrs/zustand
[5]https://reactflow.dev/
[6]https://github.com/steveruizok/perfect-freehand

the backend, adds a new node to the graph, and awaits the returned data. (4). Tool Layer (Top): This is the most visible layer. It contains operation tools and canvas control tools.

*Cluster Module.* The middle module displays selected information, including objects, images, text, and annotations. Users filter this content from the node-edge graph by selecting specific objects. A button allows users to expand or collapse this entire area.

*Text Editor Module.* The left-side text editor uses the React-Quill[7] component. It provides lightweight rich-text editing capabilities designed to align with the narrative structure.

*4.2.2 Back-End Multi-Agent Flow.* The backend is organized as a multi-agent pipeline, where each prompt-specialized LLM agent is organized sequentially. Rather than relying on a single model, the system decomposes the workflow into three cooperating functional agents:

*Narrative Construction Agent.* This agent takes multimodal context—such as canvas screenshots which incorporate all information (sketches, text inputs, and images) contained within the user-selected region to produce structured textual outputs—including user intentions and story segments. Its prompts enforce consistency with contextual information such as the existing story content in the text editor and the global stylistic constraints. In essence, the agent transforms the user's multimodal inputs into coherent narrative storylines. During implementation, the GPT-o4-mini[8] is used to process canvas screenshots and contextual information to infer user intentions because of its rapid inference speed and strong reasoning ability. GPT-4o[9] is used to refine these inferred intentions from o4-mini and contextual information into polished, coherent story segments. Prompts for precise description generation of o4-mini are shown in Appendix A.5). When applying instrumental operations for editing, the story segment from the previous card is also read in prompts and modified accordingly.

*Visual Synthesis Agent.* Using the narrative produced by the Narrative Construction Agent as prompts, this agent supports image generation using either the GPT-4o API or the FLUX diffusion model[10]. When reference images or screenshots are available—such as during instrumental-operation edits that modify images or when multimodal inputs include a base image or sketches—GPT-4o is used for image generation, leveraging its strong capabilities in image understanding and re-generating based on base images. In all other cases, the FLUX diffusion model is used due to its faster generation speed.

*Memory Agent.* The backend maintains a persistent, globally accessible memory of all previously generated image-text pairs. This agent coordinates read/write operations to this store, enabling multi-turn reuse, cross-scene integration, and contextual continuity. When the user later requests to merge scenes, change details, or perform local edits, this agent retrieves relevant image-text pairs and passes them back to the earlier agents to re-initiate the pipeline.

## 5 User Study

To evaluate the usability of Vistoria and to understand how these multimodal interactions design support creativity, we conducted an exploratory lab user study with 12 participants. We structure our user study around two complementary components:

(1) A usability evaluation of Vistoria aiming to answer the following research questions:

***RQ1:** How useful are the multimodal co-editing functions, and in what ways do they influence participants' workflows?*

[7]https://github.com/zenoamaro/react-quill
[8]We used OpenAI's o4-mini model via the OpenAI API (model ID `o4-mini`) in August 2025 [64].
[9]We used OpenAI's GPT-4o model via the OpenAI API (model ID `gpt-4o`) in August 2025 [63].
[10]We used the FLUX.1 diffusion model via the Black Forest Labs API in August 2025 [8].

***RQ2:*** *How does multimodal image–text co-editing affect participants' workload compared to a text-only baseline?*

(2) An pilot study examining the creativity support provided by Vistoria, focusing on:

***RQ3:*** *How does multimodal image–text co-editing influence participants' ideation and the development of fictional stories?*

***RQ4:*** *How does multimodal image–text co-editing influence participants' sense of agency and ownership?*

RQ1 evaluates the usefulness of these multimodality functions and the strategies participants adopted in relation to DG3 (polymorphic instrumental operations). RQ3 explores how DG1 (reifying intentions through multimodal input) and DG2 (image-text alignment) support creative ideation and narrative development. The design of the cluster panel corresponding to DG4 functions mainly as auxiliary support and is not central to our user evaluation of multimodal interaction.

Note that we used the sense of ownership to refer to the writer's "sense of possession" over the resulting narratives within the system, even the AI-generated artifacts [26, 34, 44, 84]. The declaration of the sense of agency, on the other hand, refers to the writer's awareness of "initiating, executing, and controlling" key actions in the writing and artifact editing process [44, 45, 54, 60].

### 5.1 Participants

As the system targets intermediate to expert participants who understand narrative structure and already use LLMs in their creative workflows, we recruited 12 participants (6 males, 6 females, aged 21–32, M=25.5), all with prior creative writing experience. Participants were also recruited through student organizations by sharing our study announcement in their group chats. All held Bachelor's degrees, with backgrounds spanning science, arts, design, or communication. Their creative practices included fiction writing, screenwriting, songwriting, advertising, research, and philosophy. Participants' creative writing experience ranged from under one year (n=5) to over seven years (n=1), with others reporting 1–3 years (n=2) or 4–7 years (n=3). All participants were familiar with LLMs (e.g., ChatGPT, Gemini, Claude) and had used them for idea generation, editing, descriptive support, content expansion, world-building, and style imitation. Among the 12 participants, 6 use LLMs daily, while the remaining 6 are evenly split across several times a week, occasionally, and rarely (2 in each). Each participant received a $40 USD compensation after finishing the experiment.

### 5.2 Procedure

*5.2.1 Apparatus.* Sessions were conducted on a laptop computer with keyboard and mouse for typing, dragging, and selecting. To support sketch input, we provided an external tablet (iPad) for freehand drawing on the canvas.

The baseline condition presented a side-by-side interface with a text editor and GPT-4o [62] conversational panel, enabling both manual editing and LLM-assisted text/image generation. Participants completed two story-writing tasks (Appendix A.3), each extending a given story beginning into a 300–500 word draft. Tasks were counterbalanced across conditions (Baseline vs. Vistoria).

*5.2.2 Study Procedure.* The study followed a within-subjects design with counterbalanced condition order. After informed consent and a demographic survey, participants were introduced to Vistoria through a written guide and tutorial video, followed by a short hands-on exploration (15 minutes).

In each condition, participants first focused on world-building and idea exploration (20 min) and then on refining and improving the story (20 min). We divided the writing task into two phases (exploration and refinement) to prevent participants from prematurely committing to a single storyline and to reduce fixation, thereby encouraging broader ideation before focused improvement [75].

After each condition, participants completed surveys including NASA-TLX [33] and Creativity Support Index (CSI) [15]. These surveys were chosen to be consistent with the standard measures employed in previous HCI system work on multimodal LLM-assisted ideation and storytelling [16, 23, 76]. All surveys used 7-point Likert scales. After both conditions finished, participants also completed a 15–20 minute semi-structured interview. All sessions were video-recorded via Zoom. We collected system logs, final story drafts, canvas artifacts, image–text pairs, and interview transcripts for analysis.

*5.2.3 Data analysis.* We employed a mixed-methods approach to systematically analyze three types of data.

For the qualitative interview data, we conducted an inductive, grounded theory-informed analysis [77]. First, two authors independently performed open-coding on 33% of the data, generating an initial set of 30 distinct codes. The coders then met to compare code applications, resolve discrepancies through negotiated agreement, and refine the wording and boundaries of each code. Through several rounds of discussion, they reached full consensus on all coded segments and consolidated the initial codes into a shared codebook. Using the refined codebook, the two authors independently coded half of the remaining transcripts, meeting regularly to prevent coding drift and to determine whether newly emerging codes should be incorporated. Ongoing constant comparison within and across interviews was used to further refine relationships between codes. Finally, we clustered the codes into four higher-level themes that map onto our design goals. The final codebook is shown in Appendix A.6.

Second, interaction data were analyzed through structured video coding by two authors to quantify tool usage frequency and modality switching events, and researchers aligned them with system logs on an event-by-event basis. Finally, for survey measures, we conducted paired-sample t-tests under the assumptions of normality and homogeneity of variance. When assumptions were violated, we used the Wilcoxon signed-rank test. Given the sample size limitations, we treat these quantitative results as descriptive signals intended to triangulate with and support the qualitative themes.

## 6 Study Results

### 6.1 The Usability of Vistoria

To address RQ1 and RQ2, we analyzed how participants engaged with the designed instrumental operations (*lasso, collage, perspective shift, and filter*) and the usage patterns. In addition, we incorporated quantitative survey results with qualitative data to assess how Vistoria affected participants' workload.

#### *6.1.1 The Usefulness of Instrumental Operations.*

*Lasso as a granularity controller for local-to-global rewriting.* The *Lasso* instrument is valued for enabling participants to zoom between different narrative scales. P8 emphasized, *"You can write in different scales, especially when you use the Lasso tool, in which you can extract out that specific detail, so [the story] generated in the card is more heterogeneous on the specific point."* (Figure 6 (c)) This reflects how the lasso instrument potentially enables a narrative "zoom" functionality, allowing participants to switch between macro-level story development and micro-level detail refinement within a single interface. Similarly, P9 and P11 described how the lasso enabled them to refine specific text and emphasize key points with more focused attention. In this way, the Lasso operates as a narrative instrument, turning macro-level edits into micro-level adjustments while preserving precision, enabling rollback, and sustaining fluent exploration.

| | | Vistoria | | Baseline | | Statistics | |
|---|---|---|---|---|---|---|---|
| | | mean | std | mean | std | p | Sig. |
| NASA-TLX | Mental | 5.16 | 1.528 | 3.167 | 1.337 | 0.0000 | ** |
| | Physical | 4.667 | 1.723 | 2.083 | 0.669 | 0.0002 | ** |
| | Temporal | 2.917 | 1.443 | 2.750 | 1.215 | 0.7723 | – |
| | Effort | 4.083 | 1.443 | 3.500 | 1.834 | 0.3388 | – |
| | Performance | 5.250 | 1.712 | 5.083 | 1.564 | 0.7986 | – |
| | Frustration | 2.750 | 1.183 | 1.750 | 0.622 | 0.0204 | * |
| Creativity Support Index | Exploration | 4.917 | 1.240 | 4.750 | 1.485 | 0.7126 | – |
| | Expressiveness | 6.083 | 0.996 | 4.333 | 1.775 | 0.0232 | * |
| | Immersion | 4.917 | 1.505 | 2.750 | 1.545 | 0.0006 | ** |
| | Enjoyment | 5.333 | 1.435 | 4.917 | 1.379 | 0.1753 | – |
| | Results Worth Effort | 5.250 | 1.357 | 5.583 | 0.793 | 0.5166 | – |
| | Collaboration | 5.500 | 0.674 | 4.583 | 1.505 | 0.0418 | * |

Table 1. Comparison of survey results: Vistoria vs. Baseline. Sig.: * $p < .05$; ** $p < .01$

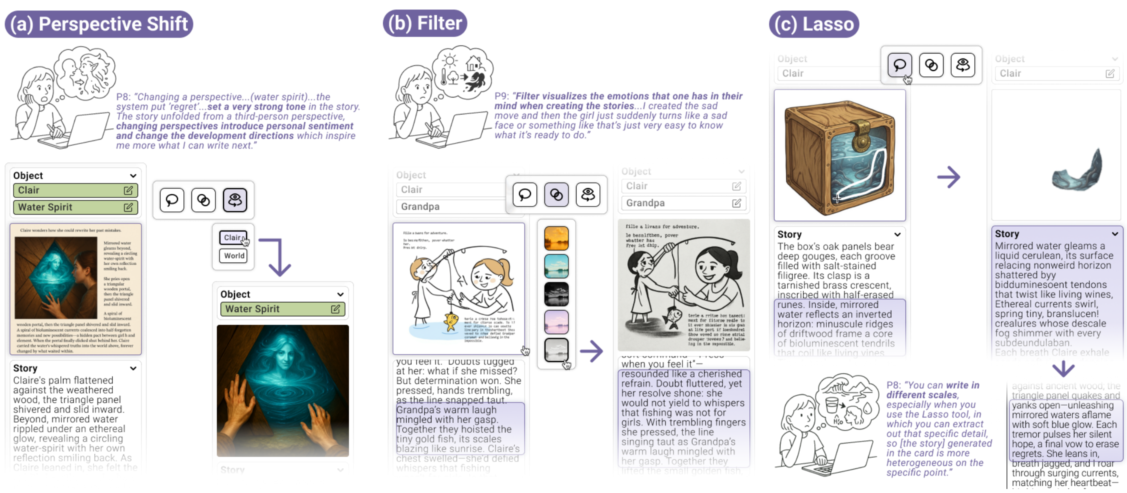



Fig. 6. (a) **Perspective Shift** changes the image viewpoint also reframe narrative voice and redirected story development; (b) **Filter** synchronizes mood and style across media—applying a visual filter also rewrites the associated text to match the intended emotional tone, ensuring stylistic coherence; (c) **Lasso** enables participants to focus at different narrative scales by extracting or isolating elements to steer local and global edits.

*Collage Function Enables Creative Recombination.* All participants used collage to merge extracted objects or scenes from different images, building connections across disparate elements. For instance, when P6 generated a scene of Maya entering a castle, he envisioned a larger structure with taller stairs. He sketched a bigger castle and mountain while expressing his intention through text, resulting in a generated card that matched his vision and was directly incorporated into his story. P11 articulated the creative freedom this technique provided: *"This technique doesn't limit me; I can create abstract or non-abstract sketches, and I can incorporate whatever I want."*(Figure 7) This multimodal recomposition enabled participants to quickly express envisioned scenes (P2) and provided "more freedom to envision and create the story" (P10). These practices highlight that collage is not merely a usability feature but a catalyst for multimodal recomposition, enabling participants to externalize, reconfigure, and expand their mental imagery into coherent narrative possibilities that text or images alone could not achieve.

*Perspective Shift as a narrative frame-shifter.* The *Perspective Shift* alters both the image and the story perspective to provide new direction for story development. P8 described how shifting perspectives changed the story direction: *"Adopting the water's viewpoint anthropomorphized the water spirit and introduced a regretful undertone that established the story's emotional framework ... changed the development direction, inspiring new writing possibilities."* P5 also experimented with this feature, incorporating a first-person voice *("I didn't expect this to be so heavy!")* adopted from the system-generated segments into her third-person story (Figure 6 (a)). Perspective Shift allowed participants to flexibly reconfigure narrative viewpoint and voice, surfacing new emotional framings and redirecting story trajectories without disrupting their ongoing writing flow.

*Filter as affective parameterization for tone alignment.* The *Filter* instrument shaped narrative emotion and tone by visually parameterizing affect. P11 noted, *"The image provides the style, which influences my story's tone and direction... Before using the filter, I can't determine tone from text alone—I need to choose between suspenseful or romantic expression ways. However, visual changes after applying a filter help me decide which feeling I want my text to have."* Similarly, P9 observed, *"Filter visualizes the emotions in my mind when creating stories... I created a sad mood with filters, and the girl suddenly turned into a sad face. It's very easy to see what it's doing and easy for me to describe later."* (Figure 6 (b)). The immediate visual feedback aligned emotional intent with text, streamlining tone-setting decisions to evaluate the usability.s.

Taken together, these instrumental operations transformed localized operations into meaningful viewpoints, scales, and tones. They appear to support participants' intended operational precision and expressiveness, while potentially reinforcing the perception–action loop during the planning and translating of story writing.

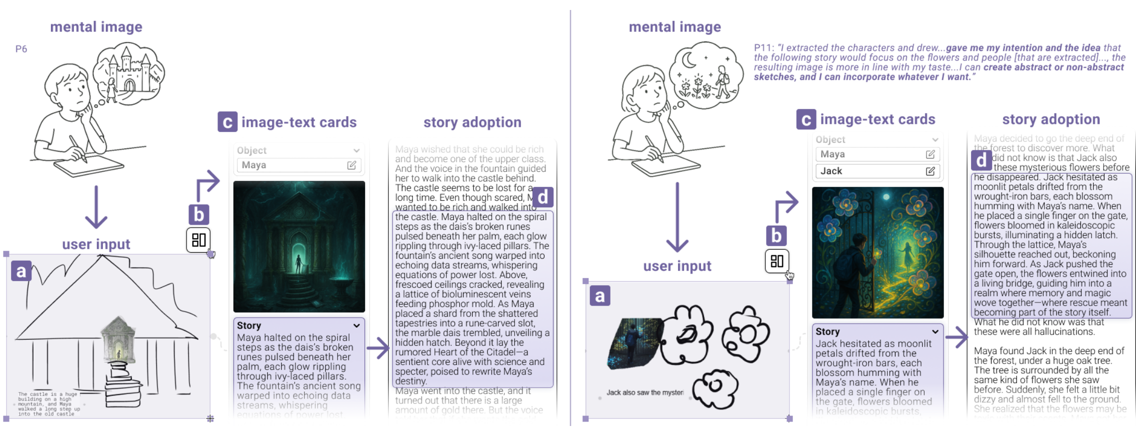



Fig. 7. Collage. Participants used Collage for creative recombination, merging extracted objects/scenes (often mixing sketches with images) to specify visualization and advance story ideas; usability was positively rated.

### 6.1.2 Tradeoffs of Multimodal Text-Image Co-editing.

*Increased Workload.* While image–text co-editing may support aspects of narrative coherence and expressiveness, our results suggest that it can also introduce additional workload. NASA–TLX scores showed significantly higher mental demand (NASA-TLX: $M_{\text{Vistoria}} = 5.16$ vs. $M_{\text{baseline}} = 3.17$, $p$=.0000; Table 1) and physical demand (NASA-TLX: $M_{\text{Vistoria}} = 4.67$ vs. $M_{\text{baseline}} = 2.08$, $p$=.0002; Table 1), plus moderately higher frustration (NASA-TLX: $M_{\text{Vistoria}} = 2.75$ vs. $M_{\text{baseline}} = 1.75$, $p$=.0204; Table 1). Qualitative analysis reveals that this increased workload reflected the higher cognitive and physical effort of coordinating across modalities and actively curating outputs compared with the GPT baseline, which involves entering text prompts and receiving output directly. Part of the mental load may have stemmed from first-time use—learning new image/text operations and switching between modalities (P1, P2, P5, P9, P10). As P2 suggested, *"The biggest burden is switching between tools to sketch or type; it takes time to learn and adapt, even though the functions are useful."* Additional difficulty also arose from unfamiliarity with the canvas interface compared with the traditional GPT interface (P3, P10).

*The Cognitive Effort of Enhanced Sense of Agency.* Nine of the participants valued the ability to maintain control of the story, and the creation process gave them a stronger sense of agency (all participants except P2, P7, and P12), and participants felt that they were directing the story's development—the final narrative emerged from their own sketches, inputs, and use of the system's instrumental operations (P4, P6, P10). However, P1, P3, P9, and P10 noted that they had to actively develop details within their own text, especially at the early stages, which could feel "a little bit frustrating" (P9). Unlike GPT, which could quickly produce long passages or propose questions to guide brainstorming (P10), the system requires participants to supply and elaborate on their own ideas before meaningful generation occurs, potentially leading to a higher mental workload. This shift demanded more cognitive and physical effort: even though sketching and annotation helped externalize mental imagery, participants noted that it felt more demanding than simply inputting and receiving GPT's ready-made text (P1, P2, P11). Thus, a stronger sense of agency may come at the cost of a higher workload.

*Validating Ideas Rather than Generating Them.* Some participants noted that the system's strengths lay in validating or expanding existing concrete mental images rather than generating new and abstract directions (P1, P11). As P1 explained, *"when I have a vague impression in my mind, I tend to generate some image-text pairs. But sometimes, once the visual appears, it fixes my imagination in a certain way in my mind, and I can no longer imagine other possibilities. In contrast, only plain text can inspire limitless imagination."* This reveals a tension: images act as concrete anchors that aid detailed development, yet their representational specificity can induce fixation by prematurely crystallizing fuzzy concepts and narrowing exploration. Similarly, P10 noted that the baseline GPT condition was superior at breaking down initial story points and directly provide additional suggestions and directions, whereas the Vistoria system primarily served to elaborate or diverge from existing visions the user already has. This suggests this multimodal approach may be valuable for participants with partially formed concepts, though potentially less helpful during the open-ended phases of ideation when abstract exploration is more important than visual specificity.

*Externalization Frees Cognitive Space.* Although participants reported experiencing higher mental workload, several accounts suggest that multimodality may have supported a more efficient allocation of attention. According to participants, the image–text pairs helped externalize fleeting ideas, preserved sensory and spatial detail, and reduced information loss when translating imagination into concrete artifacts (P4, P7). As P7 described, "By highlighting and collaging, I externalized formed ideas into image–text pairs, clearing mental space to pre-plan the next line and concentrate on the next plot beat." Taken together, these accounts indicate that while Vistoria demands more decision-making effort, it also enables a degree of cognitive offloading that shifted attention away from low-level memory maintenance toward higher-level creative synthesis. With greater familiarity, such offloading could potentially yield efficiency benefits that offset the initial overhead.

### 6.2 Creativity Support of Vistoria

To address RQ3 and RQ4, we examine how Vistoria supports intention expression through multimodal input, facilitates the ideation process through divergent exploration, as well as how image–text alignment design contributes to narrative development. We also describe how this workflow potentially enhances participants' sense of agency and ownership.

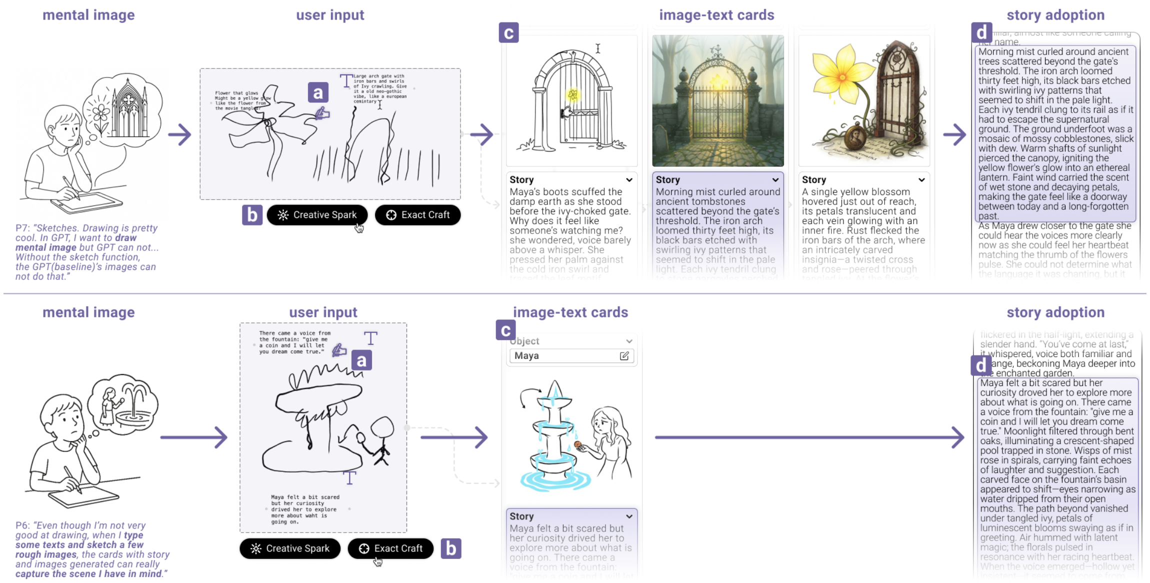



Fig. 8. Multimodal Expression. Example where sketch structure and textual details jointly yield multiple relevant images; participants valued sketches for visualizing spatial layout beyond what text-only tools could provide.

*6.2.1 Enhancing Intent Expression Through Multimodal Input.* As shown in Appendix Figure 12 (b), text served as the primary medium for generation but was consistently supplemented with sketches and images to provide more spatial information. Eight participants (P1, P2, P3, P6, P7, P9, P10, P11) stressed that combining sketches with text and images aligned outputs more closely with their creative intentions, yielding significantly higher expressiveness ratings than the baseline ($M_{\text{Vistoria}} = 6.08$ vs. $M_{\text{Baseline}} = 4.33$, $p = .023$; Table 1).

P6 captured this benefit: *"Even though I'm not very good at drawing, when I type some texts and sketch a few rough images, the cards with story and images generated can really capture the scene I have in mind."*. P7 illustrated this with a concrete case: she sketched a rough flower and archway, then added text specifying a glowing flower and a Gothic gate. The system fused the spatial layout from the sketch with textual details to generate multiple fitting images. She highlighted the unique value of sketching: *"Drawing is pretty cool. In GPT, I want to draw a mental image, but GPT cannot... the geometry of GPT-generated image is always different from what's in my head."* (Figure 8). This suggests that multimodal expression enabled participants to externalize their mental imagery and refine it into concrete, shareable representations, bridging the gap between vague internal visions and precise outputs.

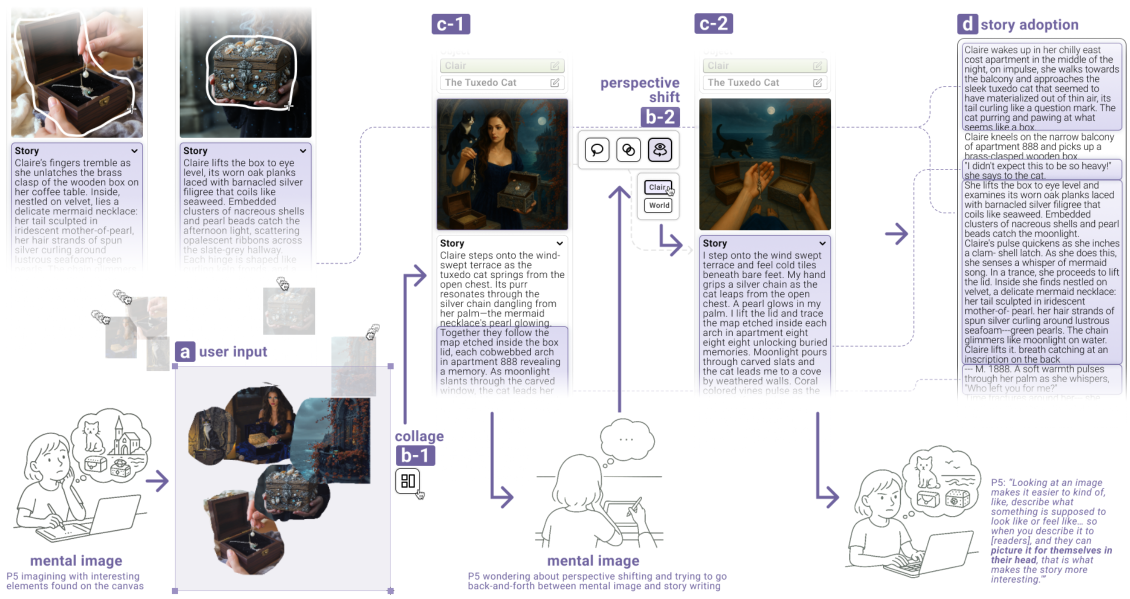



Fig. 9. Images act as cognitive scaffolds, helping participants describe unfamiliar actions or contexts more concretely. P5 constructed her final writing by referring to both images and adopting and editing related text.

*6.2.2 Bottom-up Creation to Support Divergent Exploration.* As shown in Appendix Figure 11 and Appendix Table 3, participants demonstrated greater breadth and depth of exploration using Vistoria and produced visibly more divergent narrative structures compared to the linear outputs of the GPT baseline.

The Vistoria canvas functioned as an exploratory space where participants pursued multiple storylines in parallel without linear constraints. The exploratory nature of Vistoria was also described as playful and enjoyable (P1, P7). As P1 reflected: *"Using the tool feels more like doing collage or drawing on a whiteboard or a large sheet of paper, where you can do almost anything—it's very free and interesting."* This experiential quality aligns with the higher immersion reported for Vistoria compared to the baseline (CSI: $M_{\text{Vistoria}} = 4.92$ vs. $M_{\text{Baseline}} = 2.75$, $p = .0006$; Table 1).

Rather than committing immediately to single narratives, participants typically generated multiple alternatives in early phases, positioning the system as an expressive medium rather than merely a text generator (P7, P4, P5, P8). This exploratory approach helped participants avoid early fixation and sustained creative engagement. As P8 observed: *"GPT workflow is more streamlined... top-down. Using the system feels more bottom-up. You are open to possibilities, and then you choose one way to go deep, so there's not a finite result and more possibilities being explored."*

*6.2.3 Leveraging Image-Text Alignment for Rich Descriptions and Story Progression.* When text and images appear together, the two modalities reinforce one another, enabling unfamiliar or imagined elements to be both visualized and verbalized. This cross-modal grounding supports the production of more detailed and enriched descriptions. Visual references, in particular, helped participants imagine actions or settings beyond their lived experience. P5 also noted, *"Looking at an image makes it easier to kind of, like, describe what something is supposed to look like or feel like... so when you describe it to [readers], and they can picture it for themselves in their head, that is what makes the story more interesting."* From observation, P5 closely described the scene with the cliff, the necklace, and the cat in the image in her writing and directly adopted some text generated, integrating them into her final story (Figure 9). Using the generated text and image helped produce more concrete, detailed narratives. For example, P1 used visual cues from an image showing Claire touching a letter and adopted the descriptions in the text, such as *"Claire steadies the box on her lap"* and created the narrative *"She runs her fingers over the letters, heartbeats echoing in her ears."* to form his final story (Figure 10). Here, the text description with visualization allowed P1 to capture more dynamic, sensory-rich narrative moments.

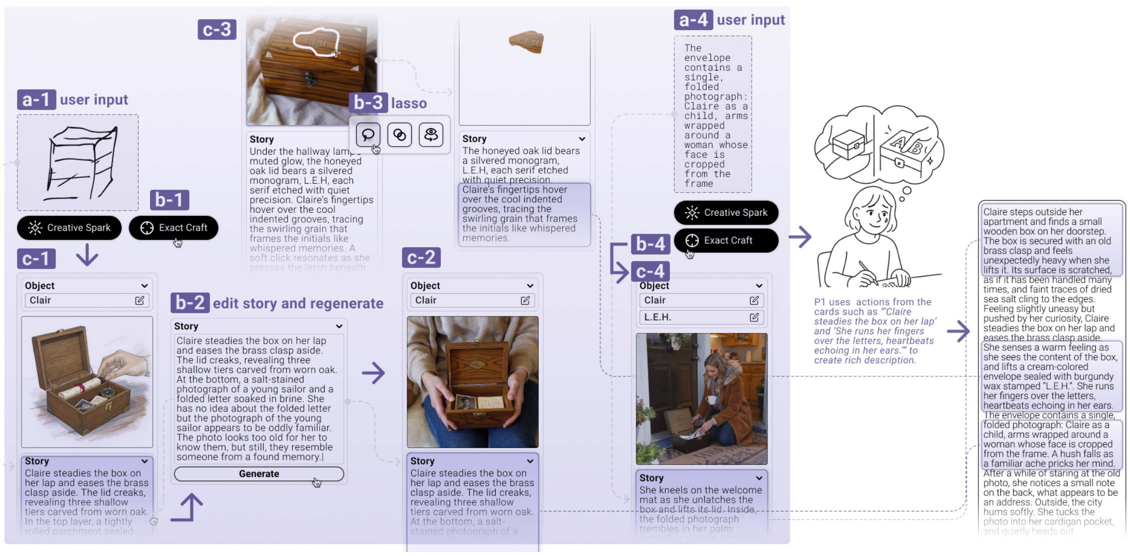



Fig. 10. Visual-to-text translation: participants turn visual cues into vivid prose that readers can picture; sample lines show how P1 used sensory-rich descriptions derived from an image-text pair.

Furthermore, unlike traditional workflows where participants must read through text to review their progress, these image-text pairs also facilitate easier tracking of story development and reduce idea drift or loss (P12, P11, P2). P12 noted that visuals alongside text helped maintain the mood and recall earlier ideas. P2 further described how the visual sequence supported planning: *"Using the system, we started with an initial image and then came up with another image… seeing the visual sequence made it easy to trace the development of the story. If you get too far down that chain and don't like it, you can just delete that node and go in a different direction. I liked visually being able to see the progress from generation to generation."* Having images presented alongside text also helped participants manage the development of their narratives and better understand where the story should go next.

*6.2.4 Preserving Sense of Agency and Ownership.* Unlike the baseline, where participants often felt like they were editors of GPT-generated text, Vistoria supported exploratory editing while preserving a sense of agency. Participants felt that Vistoria "having more sense of mastery over the content," in contrast to the baseline (P4, P6, P9). P7 felt that when using Vistoria, she was actively cutting, combining, filtering, and directing the story's trajectory, whereas with GPT, she was mostly receiving and editing what the model produced. P9 explicitly emphasized a heightened sense of agency, noting: *"This tool is more 'me'... I control characters and plots."*

Dissatisfied the baseline condition, in which GPT produced most of the content and left participants primarily in the role of adopting it (P4, P6, P7), P10 also characterized Vistoria as a supportive co-pilot rather than a substitute for their own work. Quantitative results are consistent with these perceptions: participants rated Vistoria as providing a stronger collaborative experience than the baseline (CSI: $M_{\text{Vistoria}} = 5.50$ vs. $M_{\text{Baseline}} = 4.58$, $p = .0418$; Table 1), which participants interpreted as a 'co-pilot' relationship that preserved their sense of agency.

This preserved sense of agency also led to the preserved sense of ownership. GPT outputs in the baseline condition were repeatedly described as "surface-level" (P2, P5, P7) and as "someone else's work" (P3), whereas most participants reported a stronger sense of ownership with Vistoria (P1, P2, P3, P4, P5, P6, P8, P9, P10, P11). As P5 explained: *"When using Vistoria, every idea originated from my own imagination, and the final story was formed by manipulating and combining these different self-generated ideas. This gave me a strong sense that the story was truly my own creation."*

Together, these findings suggest a shift from passively adopting model suggestions to actively creating and curating one's own generative outputs.

## 7 Discussion

### 7.1 Multimodal Instrumental Interaction

Instrumental Interaction [6] conceptualizes instruments as mediators that translate writers' actions into operations on domain objects. We operationalize this principle by reifying a set of multimodal instrumental operations (Collage, Lasso, Filter, and Perspective Shift) that simultaneously act upon both text and image narrative materials. Rather than treating images and text as separate interface elements, these instruments serve as unified interactional units that enable writers to zoom between narrative scales, reorganize multimodal story fragments, and explore divergent directions within a shared representational space to edit image and text simultaneously.

From the perspective of *Designing Interaction, not Interfaces* [7], our design moves away from adding more interface widgets and focuses instead on shaping the quality of writers' ongoing activity. Beaudouin-Lafon argues that transformative interfaces must shift attention from surface-level user-interface (UI) components toward the underlying interactional structures that support creative work [7]. Following this perspective and Dual Coding Theory's suggestions for verbal and nonverbal perception alignment [20], Vistoria's design prioritizes fluid transitions between modalities, persistent manipulable artifacts, and an iterative loop in which narrative ideas and multimodal materials co-evolve. This interaction-centered framing explains why writers perceived the system as increasing expressiveness.

Although participants generally found the instruments effective, they also reported a substantial learning curve. For novice system users, the system imposed considerable mental and physical workload. These insights suggest future design opportunities to lower workload and adapt to writers' evolving needs. In the early stages of system use, writers could express their intentions through natural language, allowing the system to suggest the most appropriate functions on their behalf [10]. As writers become more familiar with the system, they can choose functions by themselves and even define customized functions that better fit their evolving needs. This approach aligns with the emphasis on supporting diverse, situated practices rather than enforcing a fixed interface vocabulary [7]. Together, these directions point toward a design space where multimodal creativity systems integrate explicit instruments with adaptive, activity-centered interaction models to better support real-world writing workflows.

### 7.2 Designing Mixed-Initiative Multimodal Workflows

In the traditional turn-based GPT workflow, users often occupy a relatively passive or evaluative role: they receive model output and act primarily as examiners who check, accept, or correct the result [41]. In our canvas setting, users actively manipulate elements on the surface, decide which multimodal materials to combine, and select which operations to apply. From the perspective of the participants, Vistoria is not experienced as a detached “answer engine,” but as a co-pilot collaborator. Users perceive themselves as those who decide how to create, what to keep, and which tools to invoke, preserving the sense of agency and ownership.

However, this also brings cognitive effort. Precisely because the workflow is instrument and manipulation-driven, it demands that writers have a clearer sense of what they need, or at least which direction they wish to explore. When writers do not yet know what they want, the system requires them to specify intentions, choose operations that tend to lead to higher cognitive load. In contrast, a text-only GPT chat enables the rapid generation of a large amount of content, allowing for subsequent refinement through iterative prompting and selection to achieve a specific focus. Several participants, therefore, viewed our system as especially suitable when they already had some ideas or a tentative direction, rather than when they were starting from a completely blank slate.

This inspires us to design a mixed-initiative paradigm to enable smooth transitions between model-led and user-led modes to accommodate different control [35]. When writers lack a clear direction, the system should allow temporary shifts toward more GPT-like exploration, for example, generating diverse suggestions or story seeds that can then be brought back onto the canvas for instrumental refinement. It could also offer low-commitment ways to switch modalities; when intentions are clearer, it should allow more user autonomy, like fine-grained instrumental operations for precise control. Furthermore, supporting mixed-initiative involves more than simply adding a model-led mode; it also requires careful design of how and when transitions between user-led and model-led states occur. This suggests designing meta-instruments that regulate the division of labor between user and model. For example, asking the model to complete only a local fragment suggests possible next scenes based on existing user input. These mechanisms could also build on existing instrumental operations. For example, when writers use the Collage function, they can request the model to recommend collage direction, and useful elements potentially can be involved based on the existing writer’s intention to realize a true collaborative activation of ideas and shared cognition between the writer and the model. In this framing, mixed-initiative becomes an additional layer of instrumental interaction that allows users to explicitly shape who drives which parts of the creative process.

### 7.3 Limitation and Future Work

While our study provides valuable insights into multimodal story writing, several limitations constrain the generalizability and scope of our findings.

*Task Scope and Short-Term Focus.* The 300—500—word story task, while manageable for controlled evaluation, does not reflect the demands of long-form fictional writing, where authors build sustained voices, complex arcs, and intricate structures. We also did not have an opportunity to study Vistoria’s suitability for extended projects, iterative revisions, or complex narratives, leaving open questions on the consistency in long works, scalability with larger volumes, or risks of over-reliance on LLM over time of use. Moreover, evaluation relied primarily on self-reports of creativity and user experience; we did not include objective measures of story quality, originality, or literary merit.

In the future, we plan to conduct field studies that deploy this system with writers of varying expertise levels in their authentic creative contexts, observing how they integrate the tool into real writing projects over extended periods. Such longitudinal research would assess the ecological validity of Vistoria and provide insights into how writers adapt Vistoria for planning and translating across longer creative cycles [50]. We anticipate that writers might spontaneously capture inspirational moments from daily life, potentially increasing their reliance on clustering functionality as they generate more dispersed content fragments that require organization. These naturalistic studies would provide crucial insights into the tool's role in sustained creative practice, revealing usage patterns, adaptation strategies, and long-term impacts on writers' creative processes that controlled laboratory settings cannot capture.

*Participant Sample.* Our study involved only 12 participants, while this is typical for similar lab usability studies, a larger group could provide stronger statistical power, reveal more varied interaction patterns, and allow comparisons across subgroups. Furthermore, the group of participants has limited cultural and age diversity, which could have narrowed the range of narrative traditions, writing styles, and storytelling approaches represented. Future research should address these limitations by recruiting a larger and more diverse set of participants, including writers of various ages and individuals from diverse cultural backgrounds, to more fully evaluate the applicability and generalizability of the system.

*Construct Validity and Measurement Limitations.* Although we discuss constructs such as creativity, sense of ownership, and agency, these observations arise primarily from qualitative reports. Our study does not include construct-grounded measurements or comparative baselines for these phenomena. Accordingly, the interpretations should be viewed as exploratory insights rather than empirically validated effects. Future work will incorporate construct-aligned measures and validated scales such as the Mixed-Initiative CSI [42] situated within human–AI co-creativity frameworks.

*Multimodality Scope.* Our work focuses on multimodal support through text and images, but does not include other modalities. Prior research in creative writing suggests that audio can also serve as a useful medium [74], especially through nonverbal sounds and ambient effects that help shape mood and atmosphere. In future work, we plan to explore the addition of audio cues to the writing process. Such sound elements may support writers in building a stronger vibe, enhancing scene-setting, and offering an additional channel for creative inspiration.

## 8 Conclusion

This paper presents Vistoria, a multimodal image–text co-editing system that supports fictional story writing by tightly integrating image and text representations. Grounded in the WoZ co-design study, Vistoria introduces a unified set of instrumental operations (lasso, collage, perspective shift, and filter) that reify writers' intentions and enable synchronized manipulation across modalities. Through a controlled user study, we demonstrate that multimodal co-editing enhances expressiveness, immersion, and exploratory ideation. Although this multimodal workflow increases cognitive demand, participants reported preserved senses of agency and ownership, treating the system as a creative partner rather than a generative tool. We hope Vistoria highlights the opportunities for designing future writing systems that embrace multimodality as a core mechanism for ideation and narrative development.

## References


[1] Leonardo Angelini, Denis Lalanne, Elise Van den Hoven, Omar Abou Khaled, and Elena Mugellini. 2015. Move, hold and touch: a framework for tangible gesture interactive systems. *Machines* 3, 3 (2015), 173–207.

[2] Anthropic. 2024. Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet. Claude 3.5 Sonnet large language model used via https://claude.ai, June 2025..

[3] Michael A Arbib. 1992. Schema theory. *The encyclopedia of artificial intelligence* 2 (1992), 1427–1443.

[4] P Matthijs Bal and Martijn Veltkamp. 2013. How does fiction reading influence empathy? An experimental investigation on the role of emotional transportation. *PloS one* 8, 1 (2013), e55341.

[5] Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.* 59, 1 (2008), 617–645.

[6] Michel Beaudouin-Lafon. 2000. Instrumental interaction: an interaction model for designing post-WIMP user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (The Hague, The Netherlands) *(CHI '00)*. Association for Computing Machinery, New York, NY, USA, 446–453. https://doi.org/10.1145/332040.332473

[7] Michel Beaudouin-Lafon. 2004. Designing interaction, not interfaces. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (Gallipoli, Italy) *(AVI '04)*. Association for Computing Machinery, New York, NY, USA, 15–22. https://doi.org/10.1145/989863.989865

[8] Black Forest Labs. 2024. FLUX.1: Next-generation text-to-image diffusion models. https://bfl.ai/. FLUX.1 diffusion model family (e.g., FLUX.1 [pro] / [dev] / [schnell]) accessed via the Black Forest Labs API in August 2025..

[9] Christopher Booker. 2004. *The seven basic plots: Why we tell stories.* A&C Black.

[10] Samuelle Bourgault, Li-Yi Wei, Jennifer Jacobs, and Rubaiat Habib Kazi. 2025. Narrative Motion Blocks: Combining Direct Manipulation and Natural Language Interactions for Animation Creation. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS '25)*. Association for Computing Machinery, New York, NY, USA, 1366–1386. https://doi.org/10.1145/3715336.3735766

[11] Steven Brown and Eunseon Kim. 2021. The neural basis of creative production: A cross-modal ALE meta-analysis. *Open Psychology* 3, 1 (2021), 103–132.

[12] Janet Burroway, Elizabeth Stuckey-French, and Ned Stuckey-French. 2022. *Writing fiction: A guide to narrative craft.* University of Chicago Press.

[13] Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or Artifice? Large Language Models and the False Promise of Creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 30, 34 pages. https://doi.org/10.1145/3613904.3642731

[14] Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2024. Creativity Support in the Age of Large Language Models: An Empirical Study Involving Professional Writers. In *Proceedings of the 16th Conference on Creativity & Cognition* (Chicago, IL, USA) *(CC '24)*. Association for Computing Machinery, New York, NY, USA, 132–155. https://doi.org/10.1145/3635636.3656201

[15] Erin Cherry and Celine Latulipe. 2014. Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 4 (2014), 1–25.

[16] DaEun Choi, Sumin Hong, Jeongeon Park, John Joon Young Chung, and Juho Kim. 2024. CreativeConnect: Supporting Reference Recombination for Graphic Design Ideation with Generative AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1055, 25 pages. https://doi.org/10.1145/3613904.3642794

[17] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching Stories with Generative Pretrained Language Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 209, 19 pages. https://doi.org/10.1145/3491102.3501819

[18] John Joon Young Chung and Max Kreminski. 2024. Patchview: LLM-powered Worldbuilding with Generative Dust and Magnet Visualization. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 77, 19 pages. https://doi.org/10.1145/3654777.3676352

[19] John Joon Young Chung, Melissa Roemmele, and Max Kreminski. 2025. Toyteller: AI-powered Visual Storytelling Through Toy-Playing with Character Symbols. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 331, 23 pages. https://doi.org/10.1145/3706598.3713435

[20] James M Clark and Allan Paivio. 1991. Dual coding theory and education. *Educational psychology review* 3, 3 (1991), 149–210.

[21] N. Dahlbäck, A. Jönsson, and L. Ahrenberg. 1993. Wizard of Oz studies — why and how. *Know.-Based Syst.* 6, 4 (Dec. 1993), 258–266. https://doi.org/10.1016/0950-7051(93)90017-N

[22] Hai Dang, Frederik Brudy, George Fitzmaurice, and Fraser Anderson. 2023. WorldSmith: Iterative and Expressive Prompting for World Building with a Generative AI. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 63, 17 pages. https://doi.org/10.1145/3586183.3606772

[23] Ali Darejeh, Nadine Marcusa, Gelareh Mohammadi, and John Sweller. 2024. A critical analysis of cognitive load measurement methods for evaluating the usability of different types of interfaces: guidelines and framework for Human-Computer Interaction. arXiv:2402.11820 [cs.HC] https://arxiv.org/abs/2402.11820

[24] Anil R Doshi and Oliver P Hauser. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science advances* 10, 28 (2024), eadn5290.

[25] Charlotte L Doyle. 1998. The writer tells: The creative process in the writing of literary fiction. *Creativity Research Journal* 11, 1 (1998), 29–37.

[26] Fiona Draxler, Anna Werner, Florian Lehmann, Matthias Hoppe, Albrecht Schmidt, Daniel Buschek, and Robin Welsch. 2024. The AI Ghostwriter Effect: When Users do not Perceive Ownership of AI-Generated Text but Self-Declare as Authors. *ACM Transactions on Computer-Human Interaction* 31, 2 (April 2024), 1–40. https://doi.org/10.1145/3637875

[27] Inc. Figma. 2025. *Figma: Collaborative Interface Design Tool.* https://www.figma.com/ Homepage outlining Figma's design-, prototyping-, white-boarding, presentation tools and AI features.

[28] Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College Composition & Communication* 32, 4 (1981), 365–387.

[29] Gérard Genette. 1980. *Narrative discourse: An essay in method.* Vol. 3. Cornell University Press.

[30] Katy Gero, Alex Calderwood, Charlotte Li, and Lydia Chilton. 2022. A Design Space for Writing Support Tools Using a Cognitive Process Model of Writing. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022).* Association for Computational Linguistics, Dublin, Ireland, 11–24. https://doi.org/10.18653/v1/2022.in2writing-1.2

[31] Katy Ilonka Gero, Tao Long, and Lydia B Chilton. 2023. Social Dynamics of AI Support in Creative Writing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23).* Association for Computing Machinery, New York, NY, USA, Article 245, 15 pages. https://doi.org/10.1145/3544548.3580782

[32] Melanie C Green and Timothy C Brock. 2000. The role of transportation in the persuasiveness of public narratives. *Journal of personality and social psychology* 79, 5 (2000), 701.

[33] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology.* Vol. 52. Elsevier, 139–183.

[34] Jessica He, Stephanie Houde, and Justin D. Weisz. 2025. Which Contributions Deserve Credit? Perceptions of Attribution in Human-AI Co-Creation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25).* Association for Computing Machinery, New York, NY, USA, Article 540, 18 pages. https://doi.org/10.1145/3706598.3713522

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

[36] Wei Huang, Pengfei Yang, Ying Tang, Fan Qin, Hengjiang Li, Diwei Wu, Wei Ren, Sizhuo Wang, Jingpeng Li, Yucheng Zhu, et al. 2024. From sight to insight: A multi-task approach with the visual language decoding model. *Information Fusion* 112 (2024), 102573.

[37] Catherine Kanellopoulou, Katia Lida Kermanidis, and Andreas Giannakoulopoulos. 2019. The dual-coding and multimedia learning theories: Film subtitles as a vocabulary teaching tool. *Education Sciences* 9, 3 (2019), 210.

[38] Suzanne Keen. 2007. *Empathy and the Novel.* Oxford University Press.

[39] Robert E Kent. 2000. Conceptual knowledge markup language: An introduction. *Netnomics* 2, 2 (2000), 139–169.

[40] David Kirsh. 2010. Thinking with external representations. *AI & society* 25, 4 (2010), 441–454.

[41] Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *arXiv preprint arXiv:2401.16745* (2024).

[42] Tomas Lawton, Francisco J Ibarrola, Dan Ventura, and Kazjon Grace. 2023. Drawing with Reframer: Emergence and Control in Co-Creativenbsp;AI. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23).* Association for Computing Machinery, New York, NY, USA, 264–277. https://doi.org/10.1145/3581641.3584095

[43] Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A. Alghamdi, Tal August, Avinash Bhat, Madiha Zahrah Choksi, Senjuti Dutta, Jin L.C. Guo, Md Naimul Hoque, Yewon Kim, Simon Knight, Seyed Parsa Neshaei, Antonette Shibani, Disha Shrivastava, Lila Shroff, Agnia Sergeyuk, Jessi Stark, Sarah Sterman, Sitong Wang, Antoine Bosselut, Daniel Buschek, Joseph Chee Chang, Sherol Chen, Max Kreminski, Joonsuk Park, Roy Pea, Eugenia Ha Rim Rho, Zejiang Shen, and Pao Siangliulue. 2024. A Design Space for Intelligent and Interactive Writing Assistants. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24).* Association for Computing Machinery, New York, NY, USA, Article 1054, 35 pages. https://doi.org/10.1145/3613904.3642697

[44] Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A. Alghamdi, Tal August, Avinash Bhat, Madiha Zahrah Choksi, Senjuti Dutta, Jin L.C. Guo, Md Naimul Hoque, Yewon Kim, Simon Knight, Seyed Parsa Neshaei, Antonette Shibani, Disha Shrivastava, Lila Shroff, Agnia Sergeyuk, Jessi Stark, Sarah Sterman, Sitong Wang, Antoine Bosselut, Daniel Buschek, Joseph Chee Chang, Sherol Chen, Max Kreminski, Joonsuk Park, Roy Pea, Eugenia Ha Rim Rho, Zejiang Shen, and Pao Siangliulue. 2024. A Design Space for Intelligent and Interactive Writing Assistants. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* ACM, Honolulu HI USA, 1–35. https://doi.org/10.1145/3613904.3642697

[45] Roberto Legaspi, Zhengqi He, and Taro Toyoizumi. 2019. Synthetic agency: sense of agency in artificial intelligence. *Current Opinion in Behavioral Sciences* 29 (Oct. 2019), 84–90. https://doi.org/10.1016/j.cobeha.2019.04.004

[46] Zhuoyan Li, Chen Liang, Jing Peng, and Ming Yin. 2024. The Value, Benefits, and Concerns of Generative AI-Powered Assistance in Writing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24).* Association for Computing Machinery, New York, NY, USA, Article 1048, 25 pages. https://doi.org/10.1145/3613904.3642625

[47] David Chuan-En Lin, Hyeonsu B. Kang, Nikolas Martelaro, Aniket Kittur, Yan-Ying Chen, and Matthew K. Hong. 2025. Inkspire: Supporting Design Exploration with Generative AI through Analogical Sketching. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25).* Association for Computing Machinery, New York, NY, USA, Article 427, 18 pages. https://doi.org/10.1145/3706598.3713397

[48] Long Ling, Xinyi Chen, Ruoyu Wen, Toby Jia-Jun Li, and RAY LC. 2024. Sketchar: Supporting Character Design and Illustration Prototyping Using Generative AI. *Proc. ACM Hum.-Comput. Interact.* 8, CHI PLAY, Article 337 (Oct. 2024), 28 pages. https://doi.org/10.1145/3677102

[49] Lydia Listyani. 2019. The Use of a Visual Image to Promote Narrative Writing Ability and Creativity. *Eurasian Journal of Educational Research* 80 (2019), 193–223.

[50] Tao Long, Sitong Wang, Émilie Fabre, Tony Wang, Anup Sathya, Jason Wu, Savvas Dimitrios Petridis, Ding Li, Tuhin Chakrabarty, Yue Jiang, Jingyi Li, Tiffany Tseng, Ken Nakagaki, Qian Yang, Nikolas Martelaro, Jeffrey V Nickerson, and Lydia B Chilton. 2025. Facilitating Longitudinal Interaction Studies of AI Systems. In *Adjunct Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST Adjunct '25)*. Association for Computing Machinery, New York, NY, USA, Article 13, 5 pages. https://doi.org/10.1145/3746058.3758469

[51] Damien Masson. 2023. *Transforming the Reading Experience of Scientific Documents with Polymorphism.* Ph. D. Dissertation. University of Waterloo.

[52] Damien Masson, Young-Ho Kim, and Fanny Chevalier. 2025. Textoshop: Interactions Inspired by Drawing Software to Facilitate Text Editing. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1087, 14 pages. https://doi.org/10.1145/3706598.3713862

[53] Damien Masson, Zixin Zhao, and Fanny Chevalier. 2025. Visual Story-Writing: Writing by Manipulating Visual Representations of Stories. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 70, 15 pages. https://doi.org/10.1145/3746059.3747758

[54] John McCarthy, Paul Sullivan, and Peter Wright. 2006. Culture, personal experience and agency. *British Journal of Social Psychology* 45, 2 (2006), 421–439. https://doi.org/10.1348/014466605X49140 arXiv:https://bpspsychub.onlinelibrary.wiley.com/doi/pdf/10.1348/014466605X49140

[55] Fanxu Meng, Haotong Yang, Yiding Wang, and Muhan Zhang. 2023. Chain of images for intuitively reasoning. *arXiv preprint arXiv:2311.09241* (2023).

[56] David S Miall and Don Kuiken. 1999. What is literariness? Three components of literary reading. *Discourse processes* 28, 2 (1999), 121–138.

[57] Midjourney, Inc. 2024. *Midjourney (Image Generation System), Version 6.1.* Text-to-image model Midjourney v6.1 (default model until June 16, 2025) used via Midjourney web/Discord service, June 2025..

[58] Miro. 2025. Miro – AI Innovation Workspace. https://miro.com/. Accessed: 2025-12-01.

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

[60] Caterina Moruzzi. 2022. Creative Agents: Rethinking Agency and Creativity in Human and Artificial Systems. *Journal of Aesthetics and Phenomenology* 9, 2 (July 2022), 245–268. https://doi.org/10.1080/20539320.2022.2150470

[61] Cut Mukramah, Faisal Mustafa, and Diana Fauzia Sari. 2023. The Effect of Picture and Text Prompts on Idea Formulation and Organization of Descriptive Text. *Indonesian Journal of English Language Teaching and Applied Linguistics* 7, 2 (2023), 325–341.

[62] OpenAI. 2024. *GPT-4o.* Large multimodal model (GPT-4o; ChatGPT model ID `chatgpt-4o-latest`) used via https://chatgpt.com, June 2025..

[63] OpenAI. 2024. GPT-4o. https://platform.openai.com/docs/models/gpt-4o. Flagship multimodal model "GPT-4o". Used via the OpenAI API with model ID `gpt-4o` (and deployment aliases such as `chatgpt-4o-latest`) in August 2025..

[64] OpenAI. 2025. o4-mini. https://platform.openai.com/docs/models. OpenAI "o4-mini" model from the o-series. Used via the OpenAI API with model ID `o4-mini` in August 2025..

[65] Kyeongman Park, Minbeom Kim, and Kyomin Jung. 2024. A character-centric creative story generation via imagination. *arXiv preprint arXiv:2409.16667* (2024).

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

[67] Anyi Rao, Jean-Peïc Chou, and Maneesh Agrawala. 2024. ScriptViz: A Visualization Tool to Aid Scriptwriting based on a Large Movie Database. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 21, 13 pages. https://doi.org/10.1145/3654777.3676402

[68] Abbas Ali Rezaee. 2011. Investigating the effect of using multiple sensory modes of glossing vocabulary items in a reading text with multimedia annotations. *English Language Teaching* (2011).

[69] Nathalie Riche, Anna Offenwanger, Frederic Gmeiner, David Brown, Hugo Romat, Michel Pahud, Nicolai Marquardt, Kori Inkpen, and Ken Hinckley. 2025. AI-Instruments: Embodying Prompts as Instruments to Abstract & Reflect Graphical Interface Commands as General-Purpose Tools. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1104, 18 pages. https://doi.org/10.1145/3706598.3714259

[70] Karl Toby Rosenberg, Rubaiat Habib Kazi, Li-Yi Wei, Haijun Xia, and Ken Perlin. 2024. DrawTalking: Building Interactive Worlds by Sketching and Speaking. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 76, 25 pages. https://doi.org/10.1145/3654777.3676334

[71] Elizabeth B-N Sanders and Pieter Jan Stappers. 2008. Co-creation and the new landscapes of design. *Co-design* 4, 1 (2008), 5–18.

[72] Xinyu Shi, Yinghou Wang, Ryan Rossi, and Jian Zhao. 2025. Brickify: Enabling Expressive Design Intent Specification through Direct Manipulation on Design Tokens. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery,

New York, NY, USA, Article 424, 20 pages. https://doi.org/10.1145/3706598.3714087

[73] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman. 2023. Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence. *ACM Trans. Comput.-Hum. Interact.* 30, 5, Article 68 (Sept. 2023), 57 pages. https://doi.org/10.1145/3511599

[74] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman. 2023. Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence. *ACM Trans. Comput.-Hum. Interact.* 30, 5 (Sept. 2023), 68:1–68:57. https://doi.org/10.1145/3511599

[75] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 644, 26 pages. https://doi.org/10.1145/3613904.3642400

[76] Sangho Suh, Michael Lai, Kevin Pu, Steven P. Dow, and Tovi Grossman. 2025. StoryEnsemble: Enabling Dynamic Exploration & Iteration in the Design Process with AI and Forward-Backward Propagation. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 203, 36 pages. https://doi.org/10.1145/3746059.3747772

[77] Robert Thornberg. 2012. Informed grounded theory. *Scandinavian journal of educational research* 56, 3 (2012), 243–259.

[78] Kirsikka Vaajakallio and Tuuli Mattelmäki. 2014. Design games in codesign: as a tool, a mindset and a structure. *CoDesign* 10, 1 (2014), 63–77.

[79] Wen-Fan Wang, Chien-Ting Lu, Nil Ponsa i Campanyà, Bing-Yu Chen, and Mike Y. Chen. 2025. AIdeation: Designing a Human-AI Collaborative Ideation System for Concept Designers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 21, 28 pages. https://doi.org/10.1145/3706598.3714148

[80] Xiyuan Wang, Yi-Fan Cao, Junjie Xiong, Sizhe Chen, Wenxuan Li, Junjie Zhang, and Quan Li. 2025. ClueCart: Supporting Game Story Interpretation and Narrative Inference from Fragmented Clues. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 410, 26 pages. https://doi.org/10.1145/3706598.3713381

[81] Yi Wang, Jieliang Luo, Adam Gaier, Evan Atherton, and Hilmar Koch. 2024. PlotMap: Automated Layout Design for Building Game Worlds. In *2024 IEEE Conference on Games (CoG)*. 1–8. https://doi.org/10.1109/CoG60054.2024.10645627

[82] Zhecheng Wang, Jiaju Ma, Eitan Grinspun, Bryan Wang, and Tovi Grossman. 2025. Script2Screen: Supporting Dialogue Scriptwriting with Interactive Audiovisual Generation. *arXiv preprint arXiv:2504.14776* (2025).

[83] Haijun Xia, Sebastian Herscher, Ken Perlin, and Daniel Wigdor. 2018. Spacetime: Enabling Fluid Individual and Collaborative Editing in Virtual Reality. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. Association for Computing Machinery, New York, NY, USA, 853–866. https://doi.org/10.1145/3242587.3242597

[84] Yuxin Xu, Mengqiu Cheng, and Anastasia Kuzminykh. 2024. What Makes It Mine? Exploring Psychological Ownership over Human-AI Co-Creations. In *Proceedings of the 50th Graphics Interface Conference* (Halifax, NS, Canada) *(GI '24)*. Association for Computing Machinery, New York, NY, USA, Article 35, 8 pages. https://doi.org/10.1145/3670947.3670974

[85] Zihan Yan, Chunxu Yang, Qihao Liang, and Xiang 'Anthony' Chen. 2023. XCreation: A Graph-based Crossmodal Generative Creativity Support Tool. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 48, 15 pages. https://doi.org/10.1145/3586183.3606826

[86] Ryan Yen, Jian Zhao, and Daniel Vogel. 2025. Code Shaping: Iterative Code Editing with Free-form AI-Interpreted Sketching. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 872, 17 pages. https://doi.org/10.1145/3706598.3713822

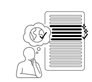
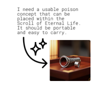
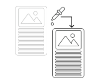
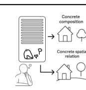
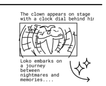
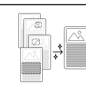
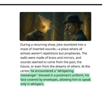
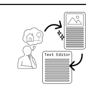
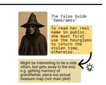
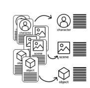
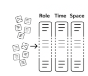
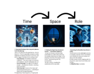
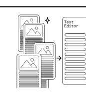
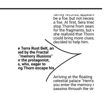
[87] Ryan Yen, Jiawen Stefanie Zhu, Sangho Suh, Haijun Xia, and Jian Zhao. 2024. CoLadder: Manipulating Code Generation via Multi-Level Blocks. , Article 11 (2024), 20 pages. https://doi.org/10.1145/3654777.3676357

[88] Zixin Zhao, Damien Masson, Young-Ho Kim, Gerald Penn, and Fanny Chevalier. 2025. Making the Write Connections: Linking Writing Support Tools with Writer Needs. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1216, 21 pages. https://doi.org/10.1145/3706598.3713161

# A Appendix

## A.1 User Behavior in the WoZ co-design study

Table 2: Results of user strategies for manipulating multimodal elements in the WoZ co-design study.

| Stage | Observed behavior | User need / Insight | N | Interaction | Example |
|---|---|---|---|---|---|
| Inputting (with Multi modalities) | **Text** | Generated prose should match the established world while allowing the injection of new elements | 83 |  |  |
| Inputting (with Multi modalities) | **Image** | Needs to inherit style/texture from references | 5 |  |  |
| Inputting (with Multi modalities) | **Sketches** | Make spatial relations/composition concrete | 8 |  |  |
| Planning | **LLM-Generated images** | Precisely locate inspiration from images | 90 |  |  |
| Planning | **LLM-Generated text** | Filter usable bits from many generations and reuse them | 90 |  |  |
| Planning/Translating | **Text notes / annotations** | Externalize temporary ideas for re-generation or final writing | 30 |  |  |
| Planning/Translating | **Collage elements** | Recompose fragments from cross-image to form new scenes | 12 |  |  |
| Planning/Translating | **Link elements** | Structure relationships between character/scene/object to connect the plot | 40 |  |  |
| Translating | **Reconfigure elements in canvas** | Reorder scattered ideas by role/time/space into place for global understanding | 4 |  |  |
| Translating | **Text (integration)** | Consolidate AI-generated text and image-inspired content into the draft | 10 |  |  |

## A.2 Behavioral interaction data gathered from participants in the user study

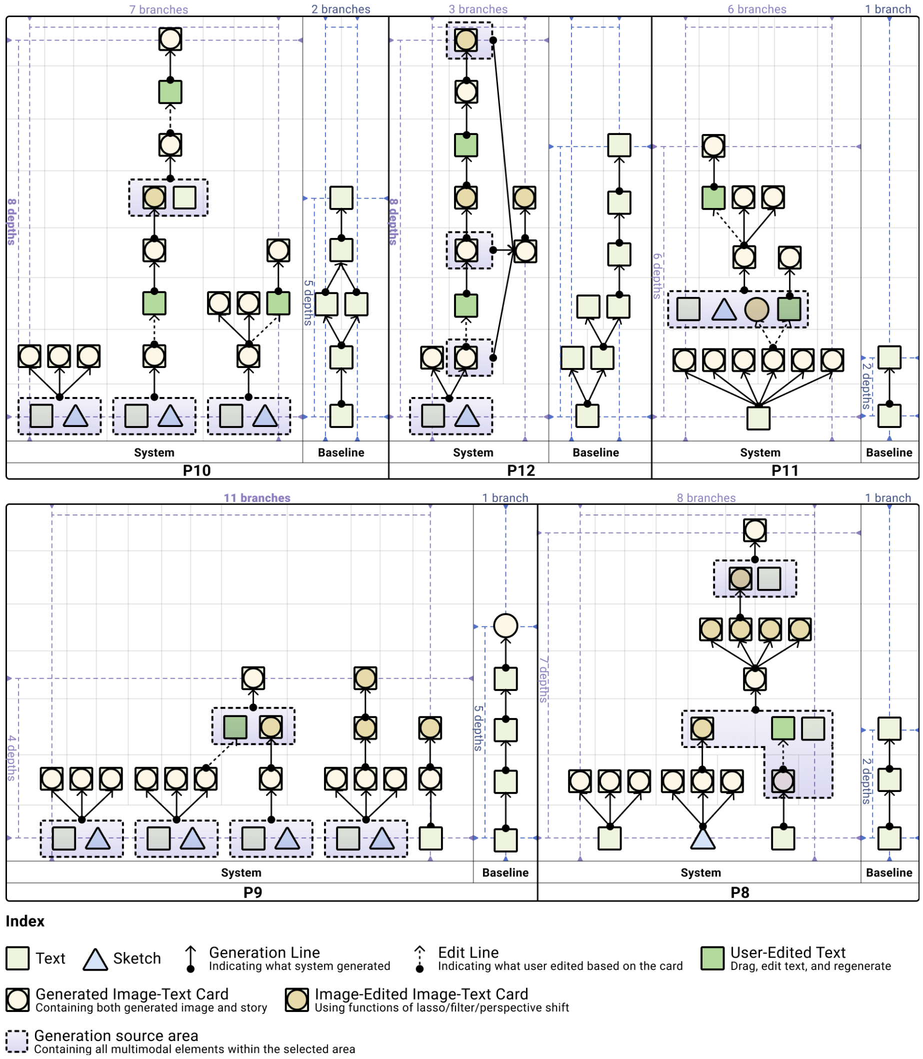



Fig. 11. Behavioral diagram contrasting Vistoria vs. baseline: participants cycle through multimodal generation, collage/recombination, and coupled image–text editing before collecting highlights for integration. Data from P8-P12 show that, compared with baseline, participants using Vistoria explored more directions, with greater divergence (Branches) within each direction. Participants also tended to pursue deeper exploration within specific directions when using Vistoria.

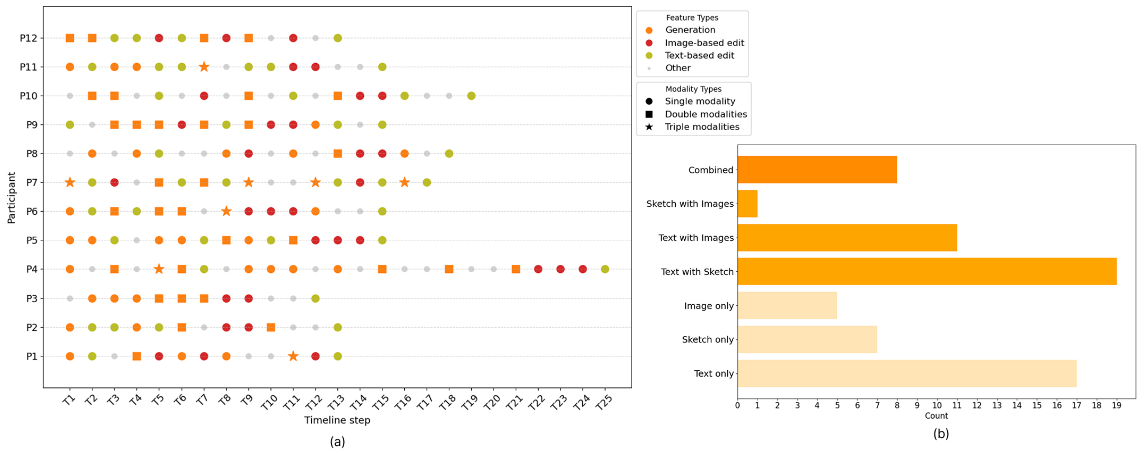



Fig. 12. Interaction records of all participants. The creative workflow begins with multimodal generation—primarily text, complemented by sketches or images to express intentions, followed by refinement and iteration using visual instruments, during which textual descriptions are continuously revised in parallel. Definition of specific behaviors: Generation includes multimodal creation of new cards using Creative Spark, Exact Craft, or Collage; image-based editing refers to operations such as Lasso, perspective shift, and filters; text-based editing covers modifications of generated story segments on the canvas as well as edits made in the text editor; other operations include updates through highlight, cluster, and upgrading global settings.

We examined the sequences of function use and compared exploration patterns across the baseline and our condition. To characterize participants' divergent–convergent behaviors during the creative process, for each task, we reconstructed exploration structures by defining Directions as top-level trajectories toward a goal, Branches as the diversity of possibilities generated within a direction, and Depth as the mean number of iterative steps within each branch to compare the exploration across the baseline and our condition.

| | Vistoria | Baseline |
|---|---|---|
| Mean # of directions | 6.92 ± 2.81 | 1.42 ± 1.08 |
| Mean # of branches | 3.00 ± 1.35 | 1.92 ± 0.67 |
| Mean depth | 1.70 ± 1.18 | 2.00 ± 1.22 |

Table 3. Descriptive statistics (mean ± SD) for Vistoria vs. Baseline. When using Vistoria, participants exhibited broader exploration; at the same time, as shown in Figure 12, they also tended to pursue individual directions with greater depth. Specifically, Directions denote the number of distinct aspects or dimensions explored when co-creation with Vistoria or baseline. Branches represent the diversity of possibilities generated within a given direction. Depth indicates the mean number of iterative steps within each branch.

### A.3 Writing Topics Used in the User Study

The two writing topic prompts used in the user study were:

Topic 1—*Claire steps outside her apartment and finds a small wooden box on her doorstep. The box is secured with an old brass clasp and feels unexpectedly heavy when she lifts it. Its surface is scratched, as if it has been handled many times, and faint traces of dried sea salt cling to the edges.*

Topic 2—*During her morning jog through the park, Maya discovers an ornate iron gate hidden behind overgrown ivy. Through the bars, she can see a path lined with luminescent flowers that pulse gently like soft heartbeats. The air carries faint whispers in a language that sounds hauntingly familiar, almost like someone calling her name.*

### A.4 Filters

The following are the types of filter supported by the *filter* instrument, showing how different types of filters are applied to image styles and mapped to text tone or emotion.

| Filter | Image Effect | Text Effect |
|---|---|---|
| **Warm** | Warm tones (gold, amber, red, orange, yellow), high exposure, strong contrast → evoke happiness, comfort, nostalgia | emphasize positivity, vitality, intimacy |
| **Calm** | Cool tones (blue, green, purple) with balanced or lower saturation → convey calmness, wisdom, introspection | Reflects contemplative and stable moods |
| **Dramatic** | Deep blacks, sharp whites, directional lighting → create intensity, mystery, urgency | Heightens stakes and emotional tension |
| **Dreamy** | Soft tones, lowered contrast, diffuse focus → suggest melancholy, intimacy, ethereality | Supports subtle, nostalgic, introspective narration |
| **Monochrome** | Removal of color, emphasis on light, shadow, texture → evoke nostalgia, timelessness, artistry | Adopts reflective and universal tone |

Table 4. Filter types with corresponding image and text effects.

### A.5 LLM Prompts

```
You are a visual story developer who analyzes screenshots to decode user visualization intent and
creates detailed story segments that bring their creative vision to life.

Process: 1) Examine the screenshot to understand what specific story content the user wants generated
by identifying: printed text({text} ) which is the primary indicator of what the user wants you to
generate, handwritten text expressing the user's desired content direction and story focus where
generation should address gaps and missing details, any images or illustrations with reference
text {previous_text} for additional context, and hand-drawn sketches representing scenes from the
user's imagination.
Synthesize these elements to understand the user's envisioned story.
The generated story should mainly focus on filling in content not covered in ({text} ) instead of
still remain unknown.
2) If hand-drawn illustrations exist in the screenshot, return the information about the story scene
conveyed in the illustration's layout; if none exist, output 'none'.
3) Generate Focused Story Content: Using the existing written passages {full_text}  only as background
context to ensure logical consistency, create a NEW detailed story segment that elaborates on a
specific scene or moment the user wants to visualize, focuses primarily on the intent expressed
in the screenshot rather than expanding the existing text, contains concrete details, character
emotions, environmental descriptions, dialogue, and interactions, maintains consistency with the
global theme {global_theme}, and can reference previous text {previous_text} if relevant to the
visualization goal.
4) The narrative should contain substantial plot or setting content, not just descriptive language.
The generated story should introducing new, insightful elements based on the context and
provide new direction of the story development that can reificate the story.
The generated stories need to be imaginative with concrete content, not filled with uncertainties.
Respond in JSON format:
{
"story": "A detailed paragraph of no larger than 100 words that creates NEW story content focused on
the user's screenshot intent, elaborating a specific scene with concrete details while
maintaining logical consistency with the background context.",
"intention": "The visualization intention read from the screenshot for story generation direction",
"sketch_information":"Regarding line sketches, integrate them with story descriptions to capture the
user's envisioned layout and scene details communicated through hand-drawn imagery, directing the image
generator to produce story scenes based on this layout guidance. Avoid generating stories
that may trigger content moderation."
}.
Provide only the JSON response without markdown formatting or additional commentary.
Let's think step by step.
```

## A.6 Codebook

**Theme 1: Instrumental Interaction**
- ⊦ **instrumental operations**
  - ⊦ granularity control (using Lasso for detail extraction)
  - ⊦ multimodal recombination (using Collage)
  - ⊦ affective alignment (using Filters for tone)
  - ⊦ perspective shift (viewpoint transformation)
  - ⊦ open new narrative direction

**Theme 2: Cognitive Process**
- ⊦ **externalization & traceability**
  - ⊦ visual history / story evolution
  - ⊦ spatial organization (grouping)
  - ⊦ visual checkpoints
- ⊦ **cognitive offloading**
  - ⊦ offloading working memory
- ⊦ **higher mental demand**
  - ⊦ not familiar with operations
  - ⊦ substantial learning effort

**Theme 3: Creative Support through Multimodality**
- ⊦ **bottom-up workflow**
  - ⊦ open-ended exploration
  - ⊦ branching storylines
  - ⊦ comparing modalities
  - ⊦ divergent exploration
- ⊦ **inspiration**
  - ⊦ serendipitous discovery (randomness as value)
  - ⊦ perspective transformation
  - ⊦ memory triggers
- ⊦ **visual–text interaction**
  - ⊦ more vivid detailed description
  - ⊦ sense of immersion

**Theme 4: Ownership and Agency**
- ⊦ **control & ownership**
  - ⊦ resisting AI takeover
  - ⊦ active curation
  - ⊦ personal style alignment
- ⊦ **metaphors of use**
  - ⊦ companion / sketchbook metaphor
  - ⊦ co-pilot
  - ⊦ free-exploration

Fig. 13. The final coding tree. Main themes are marked in bold; sub-codes represent specific strategies and behaviors observed in the study.