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

KEXUE FU*, City University of Hong Kong, China and University of Notre Dame, United States
JINGFEI HUANG*, Harvard University, United States
LONG LING*, College of Design and Innovation, Tongji University, China
SUMIN HONG, Computer Science and Engineering, University of Notre Dame, United States
YIHANG ZUO, The Hong Kong University of Science and Technology (Guangzhou), China
RAY LC, City University of Hong Kong, China
TOBY JIA-JUN LI, Department of Computer Science and Engineering, University of Notre Dame, United States

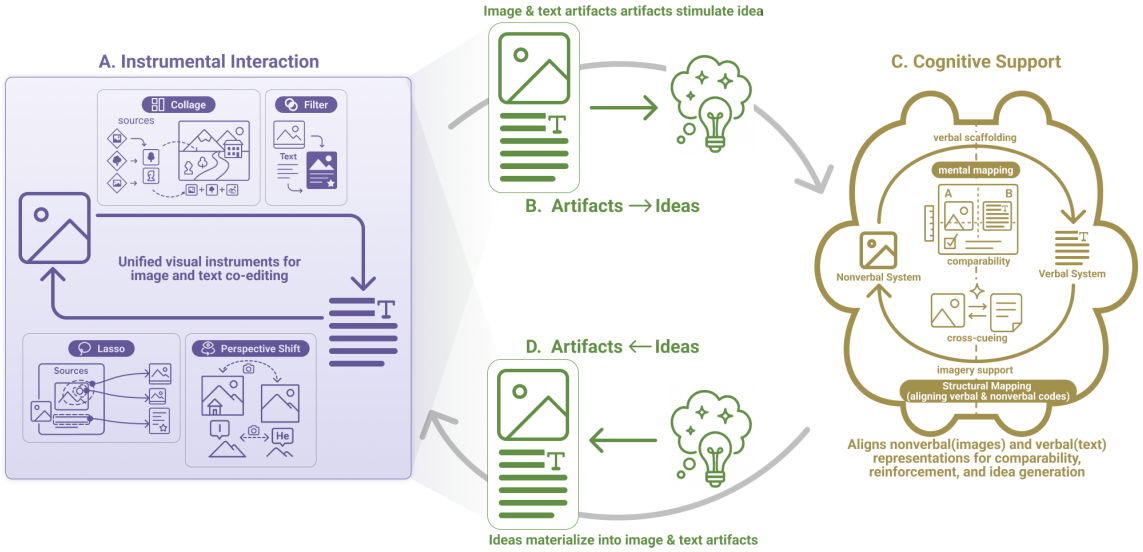

Fig. 1. Vistoria supports a cyclic workflow in which multimodal artifacts and ideas co-evolve. (A) Instrument Interaction: a unified set of tools—Collage, Lasso, Filter, and Perspective Shift—enables coupled editing of image and text to produce image–text "cards." (B) Artifacts → Idea: the resulting visual and textual artifacts stimulate new concepts and story directions. (C) Cognitive Support: a structural mapping view aligns image and text variations, making alternatives comparable and keeping narrative structure consistent. (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 authors planing and translating ideas. We present Vistoria, a system for synchronized text–image co-editing in fictional story writing that treats visuals and text as co-equal narrative materials. A formative Wizard-of-Oz

Authors' Contact Information: Kexue Fu*, City University of Hong Kong, Hong Kong SAR, China and University of Notre Dame, Notre Dame, IN, United States, kexuefu2-c@my.cityu.edu.hk; Jingfei Huang*, Harvard University, Cambridge, Massachusetts, United States, jingfeihuang@mde.harvard.edu; Long Ling*, College of Design and Innovation, Tongji University, Shanghai, China, lucyling0224@gmail.com; Sumin Hong, Computer Science and Engineering, University of Notre Dame, Indiana, United States, shong6@nd.edu; Yihang Zuo, The Hong Kong University of Science and Technology (Guangzhou), Guangdong, China, yzuo099@connect.hkust-gz.edu.cn; RAY LC, City University of Hong Kong, Hong Kong SAR, China, LC@raylc.org; Toby Jia-Jun Li, Department of Computer Science and Engineering, University of Notre Dame, Indiana, United States, toby.j.li@nd.edu.

co-design study with 10 story writers revealed how sketches, images, and annotations serve as essential instruments for ideation and organization. Drawing on theories of Instrumental Interaction and Structural Mapping, Vistoria introduces multimodal operations-lasso, collage, filters, and perspective shifts that enable seamless narrative exploration across modalities. A controlled study with 12 participants shows that co-editing enhances expressiveness, immersion, and collaboration, enabling writers to explore divergent directions, embrace serendipitous randomness, and trace evolving storylines. While multimodality increased cognitive demand, participants reported stronger senses of authorship 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 Interation, Structure Mapping, Direct Manipulation

1 Introduction

Human cognition is deeply rooted in visual processing [34, 36, 58]. We recall experiences as spatial scenes, build mental models through imagery, and rely on techniques such as memory palaces to organize abstract information. Everyday reasoning often involves manipulating internal visual representations—planning a route, rearranging objects in mind, or picturing alternative outcomes. Language use is similarly tied to imagery—comprehension frequently evokes mental pictures, and abstract ideas are often conveyed through spatial metaphors such as *path*, *framework*, or *perspective* [47]. This aligns with Dual Coding Theory, which argues that cognition integrates both verbal and nonverbal channels, with distinct representational capacities, to support imagination, reasoning, and communication [17].

Given this inherent multimodality of thought, a critical question arises: why do our most sophisticated creative practices, such as narrative writing, remain confined to unimodal tools? Fictional story writers often describe richly visualized scenes and manipulate complex spatial relationships in their imagination, yet must translate these processes through the narrow channel of sequential text [42, 51, 56, 74]. This mismatch between the multimodal nature of creative cognition and the unimodal design of creative tools constrains expressive potential [40, 82].

In fictional story writing, the most demanding phases are planning and translating [24, 28]. Authors must organize imagined scenes, characters, and events into coherent structures, then render these abstractions into linear prose [7, 16, 19, 22]. Visuals can scaffold these processes by stimulating ideas, supporting organization, and enriching narrative detail [1, 42, 51]. However, traditional visual aids have been limited in scope and scalability, offering only passive inspiration rather than dynamic, interactive support for the iterative writing process [13, 39, 42, 51, 79]. Large language models (LLMs) have unlocked unprecedented generative capabilities, enabling rapid creation of diverse narrative elements through combined text and visual imagery at previously unattainable scales [25]. LLM-powered systems have explored visual integration through character construction tools [55], visual analytics for branching storylines [50, 72], script visualization [56], diagram-prose synchronization [46], and text-editing metaphors from image manipulation [45].

However, the use of visuals remains peripheral across these approaches. They are often treated as prompts, overlays, or loosely coupled representations with unidirectional interaction. Our formative Wizard-of-Oz (WOz) co-design study with 10 writers revealed that sketches, annotations, and image manipulations are essential instruments for externalizing mental images and organizing narrative ideas. What is missing from existing LLM-powered creativity tools for fictional story writing is a unified multimodal substrate that directly supports iterative oscillation between planning and translating, where text and visuals serve as co-equal, manipulable materials in a bidirectional co-editing loop.

To bridge this gap, we developed Vistoria, a system that transforms fictional story writing from a unimodal text process into a multimodal co-editing experience where synchronized text-image manipulations through polymorphic operations—lasso, collage, perspective shifts, and filters. The design enables writers to fluidly navigate between abstract

conceptualization and concrete visualization while creating a unified multimodal substrate for narrative development. These operations enable writers to reorganize narrative elements, explore alternative viewpoints, and refine story details through a bidirectional loop where text and visuals continuously inform one another. The design of Vistoria was grounded in two key theoretical frameworks. *Instrumental Interaction* guides consistent operations across modalities to reduce switching costs and sustain flow [9, 59, 63], while *Structural Mapping* [27, 52] explains how verbal and visual representations can be aligned so edits in one modality inform the other.

We conducted a controlled study with 12 participants to evaluate Vistoria's effectiveness in supporting multimodal narrative creation. The results indicate that participants successfully integrated Vistoria's synchronized text-image co-editing into their creative workflows, employing four distinct multimodal strategies to enhance storytelling. Participants demonstrated enhanced creative expression by using images as catalysts for richer descriptions, leveraging visual editing operations to drive narrative pivots, and employing multimodal elements to externalize and interconnect story branches. The cross-modal dialogue facilitated creativity through divergent exploration of multiple narrative directions, serendipitous discoveries from AI-generated content, and culture-driven associations that activated personal memories and references. While multimodality increased cognitive load, study participants reported heightened authorship, agency, and immersion, characterizing the system as a collaborative partner that expanded their imagination. Ultimately, creativity emerged from the dynamic interplay between modalities, where text and imagery continuously shaped one another.

In summary, this work contributes:

- A WOz co-design study with 10 writers examined the practices and needs of using multimodal elements (text, sketches, and visual imagery) 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 polymorphic operations and a bidirectional editing loop to support planning and translating in fictional story writing;
- A controlled usability study with 12 participants demonstrates the effectiveness of Vistoria, showing that multimodal co-editing enhances expressiveness, coherence, and authorial agency 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 [22]. Writers must invent narrative worlds and characters while ensuring coherence, which poses unique cognitive challenges: abstract, non-verbal mental images must be transformed into structured narrative elements and then into text [7].

The Cognitive Process Model of Writing [24, 28] frames writing as recursive processes of planning, translating, and reviewing. In fictional story writing, the transition from planning to translating is especially demanding, as writers move from imaginative constructs to linear verbal representation, imposing a high cognitive load. However, visual representations can scaffold this process by externalizing abstract ideas. Research shows that picture prompts improve writing coherence [51], and visual images stimulate creativity in narrative writing [42]. In practice, sketches, maps, and diagrams help structure plot, setting, and character relationships during planning, and serve as cognitive anchors during translating to support coherence and consistency [82]. Dual Coding Theory [17] explains these benefits: verbal and imagery 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 [8].

Building on this foundation, a variety of creativity support tools have incorporated visuals into different phases of writing. Planning-focused systems such as CCI, Sketchar, and CharacterMeet assist authors in character and world development, often through image-guided backgrounds or conversational refinement of characters [43, 54, 55]. Translation-focused tools like ScriptViz and Script2Screen aim to align textual composition with visual referents, either by retrieving reference visuals from movie databases [56] or by synchronizing scriptwriting with audiovisual scene creation [74]. Complexity management systems, for example, WhatIF [50], ClueCart [72], and PlotMap [73], help writers maintain structural coherence by visualizing branched narratives, organizing narrative clues hierarchically, or integrating spatial layouts with textual plot structures.

Despite these advances, visuals in prior systems remain peripheral: they serve primarily as references, scaffolds, or analytic overlays, rather than as co-equal, manipulable materials. This treatment constrains multimodal creative expression by limiting writers to either text-first workflows or loosely coupled visual aids [82]. To bridge this gap, our work introduces a system that treats visuals and text as fully integrated, co-editable artifacts, directly supporting the **planning** and **translating** phases of **fictional story writing** so that imagination can be externalized and refined seamlessly across both modalities.

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 [45, 59]. Powered by advanced LLMs, these systems address fundamental barriers through three complementary mechanisms. First, externalizing creative structures. Tools like AI-Instruments [59] and Brickify [63] turn abstract intentions into manipulable interface objects or reusable design tokens, making spatial, stylistic, and compositional relations tangible. Second, reducing cognitive load with multimodal input: DrawTalking [60] and Code Shaping [80] distribute expression across sketch, voice, or annotations, lowering the reliance on precise verbal specification. Third, supporting iterative refinement. Inkspire [41] and AIdeation [71] accelerate variation and exploration, enabling rapid cycles of sketch-to-output or recombination of references. However, most current systems remain fragmented: they either support isolated stages (e.g., ideation, editing) or couple modalities only loosely, often within a one-shot prompting paradigm that underutilizes LLMs' potential for sustained dialogue and refinement [19].

Seeking tighter coupling for fictional story writing, recent systems push integration of multimodal interaction in different ways [15, 16]. WorldSmith grounds worldbuilding in text–sketch edits and region-based filling to support users iteratively visualize and modify elements of their fictional world [19]. XCreation [79] supports cross-modal storybook creation by integrating an interpretable entity-relation graph, improving the usability of the underlying generative structures. Toyteller [16] maps symbolic motions to character actions. Visual Writing allows writers co-edit stories by directly manipulating visual representations of entities, actions, and events alongside text, enabling easier navigation, editing, and exploration of narrative variations [46]. These systems highlight the transformative potential of LLMs to couple narrative and visual modalities, yet they still lack a unified, synchronized substrate that supports fictional story writing as a continuous, multimodal workflow. Our work advances this direction by introducing such a substrate, where LLM-generated texts and visuals act as co-editable inspirational materials for iterative, coherent story development.

2.3 Instrumental Interaction and Structural Mapping Theory

Instrumental Interaction is central to understanding how users control and refine digital systems. Beaudouin-Lafon [9] 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 [38]. 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 [81] and in Textoshop's reification of abstract text editing commands (e.g., tone adjustment, boolean operations, layers) into direct manipulation tools for text [45]. A second principle is polymorphism, where the same instrument applies across contexts [44]. This reduces cognitive load by enabling predictable, transferable patterns, e.g., copy–paste works consistently across text, images, and files [3], and scrollbars operate similarly across documents, spreadsheets, and browsers [44]. Finally, reuse allows users to replay or adapt prior operations, from macros to redo commands [59]. Systems like Spacetime exemplify this by objectifying space, time, and actions into persistent containers, enabling edits to be carried forward as manipulable entities [77]. Together, these principles reduce cognitive burden by externalizing interaction histories, making them manipulable, transferable, and extensible.

In our system, we extend this perspective to AI support for fictional story writing, where imagination must be continuously externalized and revised. We design interaction instruments that reify narrative elements into manipulable objects and, through polymorphism, let the same tools operate across LLM-generated text and images. This integration of reification and polymorphism enables writers to fluidly adjust multimodal outputs as persistent, tangible entities.

Structural Mapping Theory (SMT) is central to understanding how users interpret and compare complex representations. Gentner [27] proposed SMT to describe how people form abstractions through analogical alignment of relational structures. Rather than relying on surface similarities, SMT highlights the role of alignable differences, which reveal systematic variations and help learners construct new knowledge. Building on this foundation, SMT-informed design emphasizes preserving object structure rather than abstracting it away. At the collection level, showing complete objects allows users to perceive relationships across the dataset [76, 78]. At the part level, systems can highlight structural correspondences by surfacing identity matches (literal overlaps) [30], property matches (shared attributes) [35], and conceptual matches (semantic commonalities) [31]. These correspondences support the iterative process of aligning, re-aligning, and refining mental models. Prior visualization and learning systems demonstrate the benefits of making such correspondences explicit, between code solutions [30], textual variants [31], or large language model outputs [5, 29], helping users notice critical differences, generalize across examples, and avoid disengagement when facing "walls of text." SMT thus serves as both a cognitive theory and a design constant for interactive systems that deepen users' reasoning about collections of related objects.

In our system, SMT informs the design of synchronized text–image co-editing: textual fragments are explicitly mapped to visual elements based on their shared properties, matched concepts, or references to the same object, so edits in one modality lead to structurally aligned updates in the other, allowing users to compare, reorganize, and refine story elements across modalities.

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 [14, 15, 19, 82]. However, current tools treat images

and texts as supplementary rather than integral to the creative process, leaving unclear how writers actually integrate multiple content types.

To address this gap, we conducted a Wizard-of-Oz co-design study [18] examining how creators use multimodal content (images, text, sketches) when planning and drafting fictional stories [24, 28, 82]. 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.

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 [18, 61, 70].

3.1.1 Participants. We recruited 10 participants, each with an two or more years of creative writing experience. 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.

Experimental Setup Three days prior to 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 [10], which served as the foundation for subsequent ideation and content development. The 90-minute main session took place in Figma[23], which provided a collaborative canvas for free sketching, arranging text–image materials, and treating multimodal inputs as manipulable artifacts.

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 a simulated AI assistant operated by the researcher (wizard). Rather than working toward a fixed output, participants iteratively developed stories of about 300 words while envisioning how the tool itself should behave. They freely requested content through voice, text, or hand-drawn sketches, and called upon the assistant for image and text editing as new needs emerged. The wizard responded in situ to simulate intelligent system behaviors, encouraging naturalistic exploration. To support this process, the experimenter used Claude [4] for text generation (drawing on outlines and randomness to stimulate ideation) and Midjourney [49] and GPT-4o [53] for image generation from prompts, sketches, and editing instructions. 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.

3.2 Formative Study Findings

3.2.1 Using multimodal input to reify vague ideas. As shown in the Table 1, 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 JoJo appeared on stage and created a sketch showing "JoJo emerging from a giant spinning wheel" with textual annotations describing the intended atmosphere, hoping AI could elaborate narrative details. **Text** functioned as the primary

medium for conveying intent, allowing participants to express connections between desired content and existing stories while clearly communicating their intentions. Participants frequently used textual annotations to specify story parts or image types for AI generation and guide generation direction. **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."*

Participants most often sought AI elaboration on **characters**, **objects**, and **scenes**, expressing the need for assistance with character design refinement, setting depictions, or object visualization. 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.

These findings highlight the value of systems that accept heterogeneous multimodal input and help co-designers transform nascent ideas into concrete narrative materials.

Table 1: Results of user strategies for manipulating multimodal elements in the WOz co-design study.

Stage	Observed behavior	User need / Insight	N	Interaction	Example
Multi-modal input	**Text**	Generated prose should match the established world while allowing the injection of new elements	83		
Multi-modal input	**Image**	Needs to inherit style/texture from references	5		
Multi-modal input	**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		

Continued on next page

Table 1: Results of user strategies for manipulating multimodal elements in the WOz co-design study. (Continued)

Translating	**Text (integration)**	Consolidate AI-generated text and image-inspired content into the draft	10		

3.2.2 Text–Image Interplay as Complementary Design Moves. Participants perceived text and images as distinct yet complementary in their creative writing. We observed an oscillation between abstraction (text) and concreteness (images) as a recurring co-design pattern, which reveals design opportunities for multimodal authoring tools.

Text as open imagination. Participants described text as a "blank canvas" for associative and structural thinking. P4 noted, *"text allows me to imagine many things in my mind,"* and P1 emphasized its "infinite imagination on a blank page." P8 highlighted that text helped set up 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 prematurely 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 likened it to a "look-and-write exercise" that scaffolds scene construction. These accounts highlight the value of image outputs not just as illustrations but as provocations. Designers 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. This interplay suggests design opportunities for systems that orchestrate smooth transitions between modalities, enabling users to fluidly oscillate 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."* Such practices illustrate 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 emphasize annotation as both a vehicle for iteration and a bridge to subsequent writing.

Together, these findings suggest systems should enable flexible recombination, localized editing, and traceable annotations to help creators turn multimodal inspiration into actionable writing material.

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. 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" for ensuring cross-chapter consistency and avoiding logic conflicts (P4, P7).

These challenges point 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 [24, 28, 82].

- **DG1: Reifying Multimodal Intentions into Manipulable Artifacts.** Informed by findings in Section 3.2.1, our system should reify multimodal inputs (text, sketches, images) into first-class artifacts that can be annotated, rearranged, and recombined. Drawing on Instrumental Interaction, such reification should preserve fleeting intentions as reusable materials for later iteration; following SMT, these artifacts should function as alignable units that support comparison, linkage, and cross-modal reorganization.
- **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, image refinements can inform descriptions, and alternative versions should make differences explicit. Grounded in Dual Coding Theory and SMT, this alignment should support cycles of divergence and convergence by allowing users to oscillate between abstraction (text) and concreteness (images).
- **DG3: Enabling Polymorphic Cross-Modal Manipulation.** Informed by findings in Section 3.2.3, our system should support direct manipulation interactions that apply uniformly across modalities, guided by Instrumental Interaction's principle of polymorphism. This uniformity should reduce switching costs and empower writers to fluidly manipulate and edit textual and visual fragments while maintaining narrative coherence.
- **DG4: Organizing and Reusing Fragments into Coherent Narratives.** Informed by findings in Section 3.2.4, our system should cluster and organize fragments and fleeting ideations, surface hidden connections, and consolidate dispersed inspirations into coherent, evolving narratives. It should further support the reuse of narrative units (e.g., characters, settings, objects) to ensure consistency and foster pathways for story progression.

4 Vistoria System

In this section, we present the key features of Vistoria. As shown in Figure 2, the interface comprises three primary components: a left text editor, a central collapsible cluster panel, and a right canvas interface. The text editor houses the current story draft, serving as contextual information for content generation. The right canvas supports freeform

sketching, text input, and comprehensive editing tools. The central cluster panel aggregates highlights and annotations from canvas content, organizing them by characters, objects, and scenes for reference and summary.

Together, these components create a workflow where narrative text, visual artifacts, and organizational structures remain continuously aligned, enabling writers to externalize, manipulate, and recombine ideas and materials across modalities.

4.1 Key Features

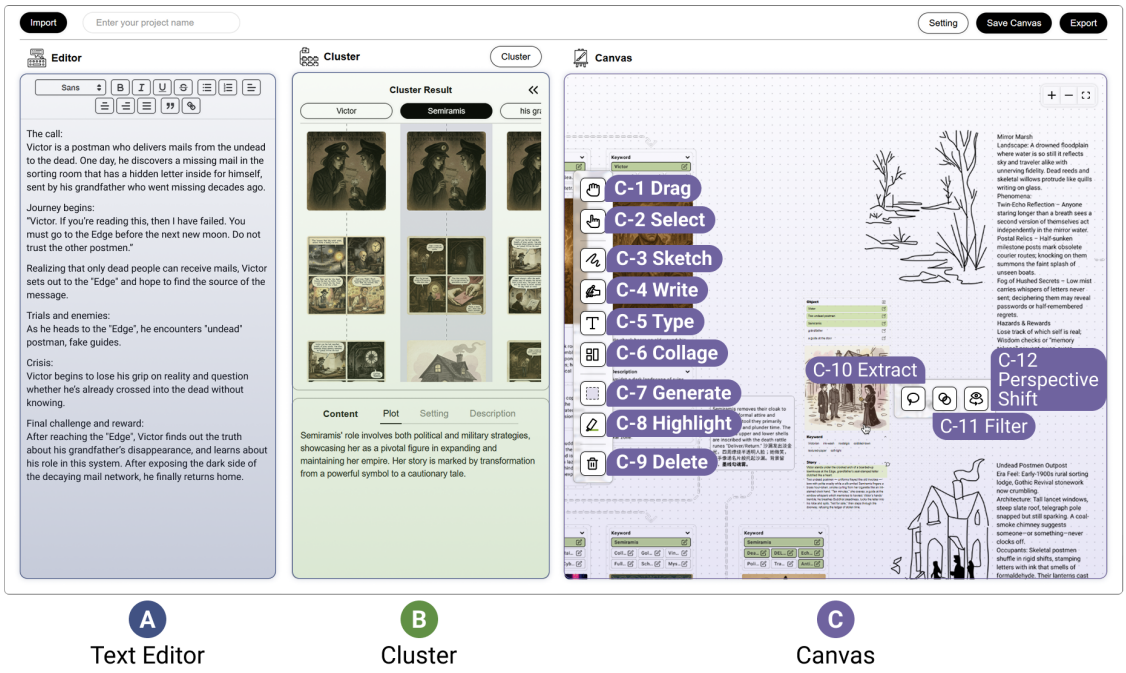

Fig. 2. Vistoria's interface: a left text editor, a central Cluster panel(can be collapsed when not used), and a right free-form Canvas. Cards (image-text pairs with main object keywords) are generated from user intentions and serve as the basic units; two modes—Exact Craft and Creative Spark—support precise realization vs. exploratory variation.

4.1.1 Reifying Intention through Multimodal Generation. Writers often externalize early intentions through sketches, short notes, or reference images, but these expressions are typically fleeting and difficult to refine into coherent narrative materials. Supporting the preservation and refinement of such multimodal intentions is therefore essential for sustaining ideation (DG1). To address this, Vistoria reifies multimodal inputs into persistent *Cards* that bundle an image, a story fragment, and extracted narrative objects (characters, objects, scenes). Drawing on the principle of *Reuse*, Vistoria reifies sketches and text from ephemeral marks into manipulable and reusable artifacts that remain on the canvas and can be invoked again as elements for regeneration, enabling writers to carry forward and iteratively develop their creative intentions. At the same time, we treat cards as *alignable units*: text and their corresponding images are designed to carry the same underlying meaning, with images serving as visualizations of the narrative described in text.

Vistoria further balances precision and exploration by offering two complementary generation modes (Figure 3). In *Exact Craft* mode, single cards closely adhere to the author's expressed intention to concretize specific ideas. In *Creative Spark* mode, three cards were generate represent diverse options based on writer's intention, the system deliberately introduces variation around characters, settings, or objects, providing alternative prompts that can inspire new directions.

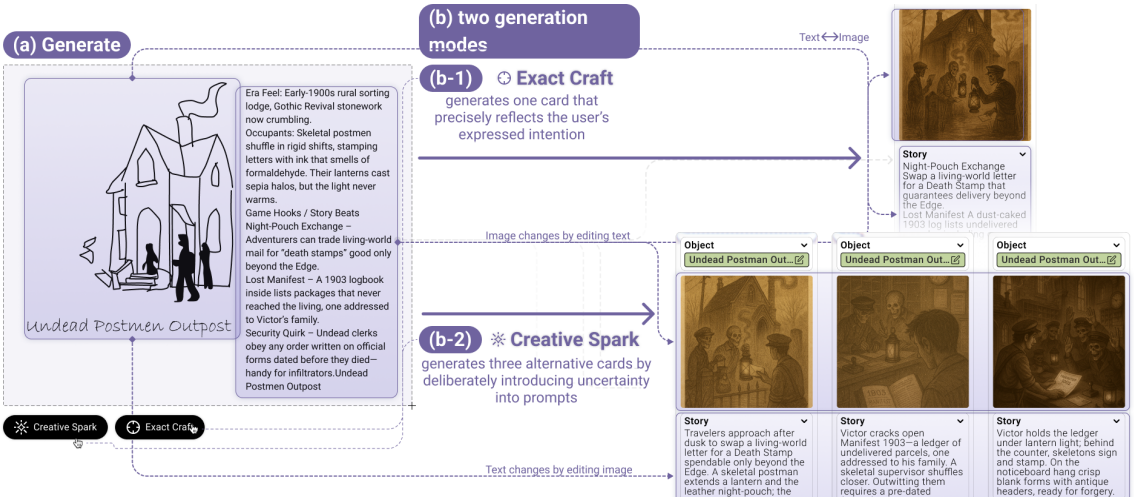

Fig. 3. Vistoria balances precision and exploration through two complementary generation modes: in Exact Craft mode, a single card closely follows the author's intention to concretize ideas; in Creative Spark mode, three cards are generated with deliberate variations in characters, settings, or objects, offering diverse options to inspire new directions.

4.1.2 Synchronized Image–Text Co-Editing through Visual Instruments. Fictional story writing benefits from fluid movement between abstract textual reasoning and concrete visual imagination, yet existing tools separate these modalities and disrupt creative flow (DG2, DG3). To address this, we introduce a set of *visual instruments* (Figure 4) designed around three principles: (1) *Reification*, which draws on familiar operations from image editing (e.g., lassoing, collaging, filtering) and transforms them into persistent, manipulable operations that apply equally to text and image elements [45, 68]; (2) *Polymorphism*, which ensures the same operations apply uniformly across text and images; and (3) *Structural Mapping*, which synchronizes changes so that visual and textual representations carry the same narrative meaning. Together, these principles reduce modality switching costs while maintaining coherence between story segments and imagery.

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 generates 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 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. Structural mapping ensures that 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)).

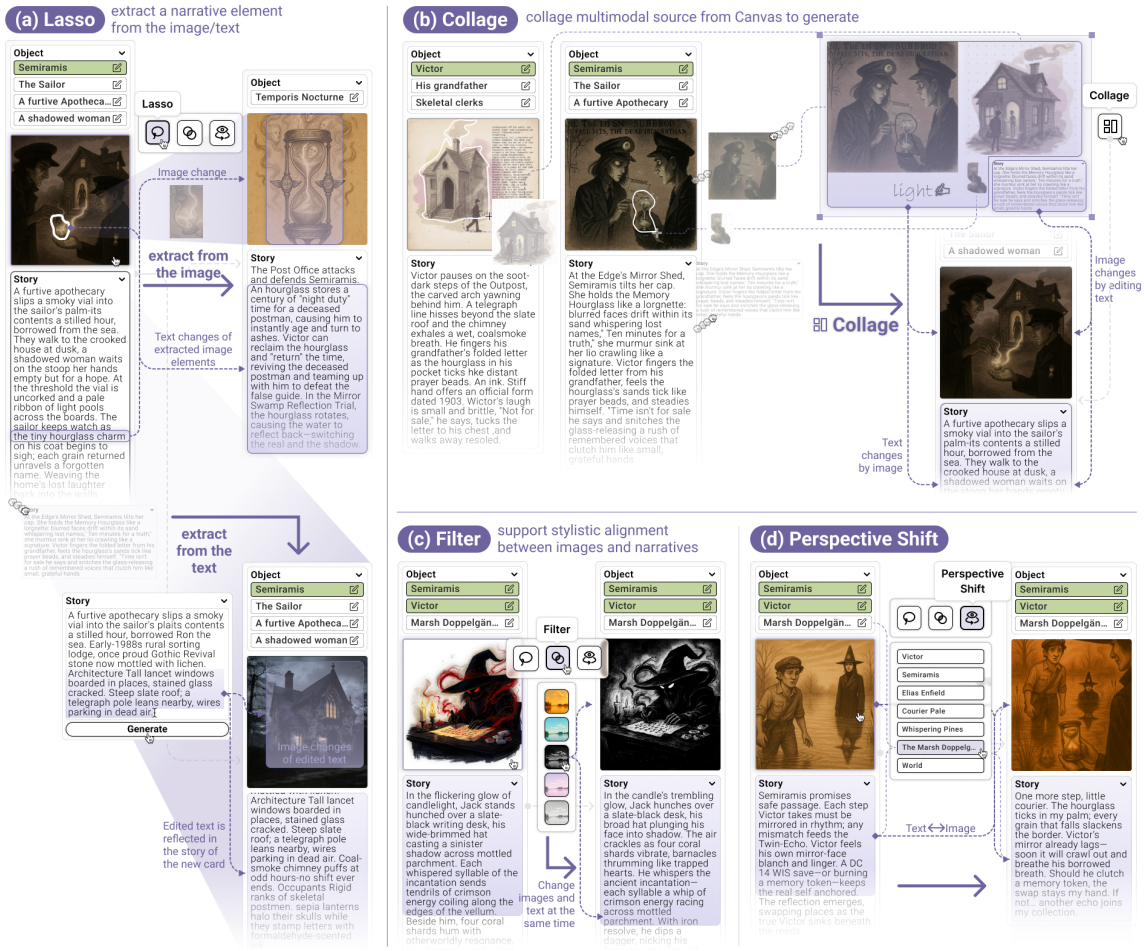

Fig. 4. A set of visual instruments for image-text co-editing to enhance planning and translating of fictional story writing: (a) Lasso selects regions for coupled image–text edits. (b) Writers extract elements and compose across cards to discover new narrative directions, using collage alongside other instruments to recombine sketches, images, and text on the canvas. (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.

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. Through *polymorphism*, 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. Guided by *structural mapping*, 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 [6], and enhance the emotional resonance of literariness [48]. In Vistoria, the *filter* instrument reifies this abstraction into a concrete tool: applying a "melancholic" or "dreamy" filter adjusts both the style of the image and the tone of the accompanying prose (Appendix Table 4). Through polymorphism and structural mapping, the same filter operation works seamlessly across modalities, leveraging the correspondence between visual style in images and emotional tone in text to act on both simultaneously. 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 shifts, as narratology highlights that changes in voice and focalization fundamentally reshape how events and characters are perceived [26]. 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 [37]. 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. *Structural mapping* ensures that the focalization 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 structural awareness in a synchronized manner (Figure 4 (d)).

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 binds 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 to manually scan scattered cards and provides writers with a dynamically updated, object-centered workspace. Selecting an object reveals its complete associated materials, such as 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 [11], enabling writers to iteratively scaffold coherent storylines from previously disjointed ideas.

4.2 Implementation

System Architecture. We adopted a decoupled front-end/back-end architecture. The React front-end enables efficient rendering for complex interactive interfaces, while the Flask back-end flexibly handles model calls with minimal overhead. Axios manages asynchronous communication between layers. The canvas module uses React-Flow for

dynamic node-edge relationships and integrates "path to SVG" for natural interactions like freehand drawing and lasso selection. A DOM-based screenshot function ensures visual consistency in exported images. The text editor integrates React-Quill for rich text editing with lightweight customizations. Global state (text content, canvas nodes, selections) is centrally managed via Zustand to ensure consistency across modules. Per-session persistence relies on `sessionStorage`, which preserves data during the study session and clears on tab close to protect privacy.

Back-end Multi-Agent flow. The back-end bridges front-end actions with language and vision models through multi-agent orchestration. Every canvas action (e.g., selection, lasso, collage, inline text) is rasterized into screenshots paired with local context (outline, prior cards, global theme). A text agent (o4-mini) examines screenshots using templated prompts from a prompt library to extract user intent, recover sketch layouts, and write text consistent with surrounding outlines. The agent returns structured JSON (story text, intent, layout hints, semantic tags) that downstream visual agents use for synthesis and alignment.

Image Generation and Editing. An image agent invokes Flux Kontext Pro with consolidated prompts that fuse screenshot-derived intent, story fragments, and style controls. Sketches serve as structural scaffolds respecting composition while avoiding duplication; reference images act as content anchors, encouraging variation over copying. When neither is present, Flux operates in free mode guided by text-agent intent and style.For editing, front-end instruments (e.g., perspective shifts, filters, collage) are translated into prompt deltas through GPT-4o's image pathway, yielding coupled visual updates. In parallel, the text agent revises tone or viewpoint for synchronized multimodal transformation. All artifacts are persisted as *cards* (image, story, objects keywords), enabling highlight-and-cluster operations that help reassemble fragments into coherent narratives.

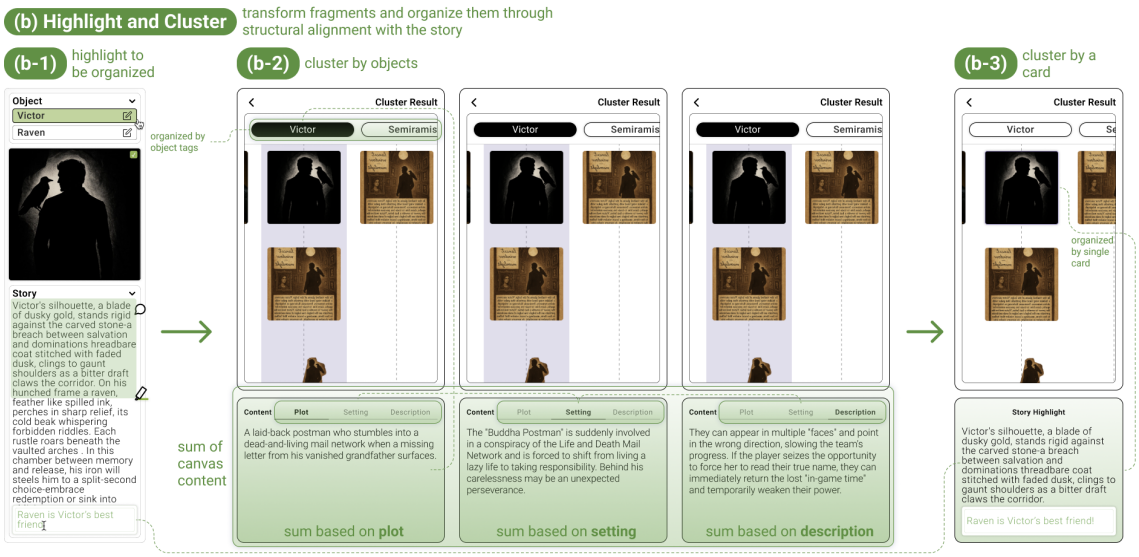

Fig. 5. Writers highlight objects and text segments on cards(b-1); the Cluster panel aggregates these by character/object/scene and can auto-summarize settings/plot/description about a certain objects to guide final writing(b-2);Clicking on a specific image reveals the corresponding highlights and comments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

5 User Study

To evaluate the usability, effectivness and usefulness of Vistoria, we conduct a lab user study with 12 participants. The study aims to answer the following research questions:

- **RQ1:** *What interaction patterns and strategies emerge when participants use a image-text co-editing process?*
- **RQ2:** *How does image-text co-editing facilitate and shape the cognitive process of story creation, in terms of ideation, iteration, and refinement?*

5.1 Participants

We recruited 12 participants (6 males, 6 females, aged 21–32, M=25.5), all with prior creative writing experience. Most held Bachelor's or Master'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 large language models (e.g., ChatGPT, Gemini, Claude) and had used them for idea generation, editing, descriptive support, content expansion, worldbuilding, and style imitation. This diversity reflects our design goal of supporting not only expert writers but also a broader range of users. Each participant received $40 USD compensation.

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 conversational panel, enabling both manual editing and LLM-assisted text/image generation. Participants completed two story-writing tasks (Appendix A.1), 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 worldbuilding 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 [67].

After each condition, participants completed surveys including NASA-TLX [33], Creativity Support Index (CSI)[12], self-perceived ML experience[13, 75], and (for Vistoria only) perceived feature usefulness. All instruments used 7-point Likert scales. Following both conditions, participants 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.

First, interview transcripts were independently open-coded by two authors following an inductive, grounded theory–informed process [69]; initial codes were iteratively refined into a shared codebook, with themes developed through constant comparison. Disagreements were resolved via negotiated agreement, and analytic memos were maintained for transparency.

Second, interaction data were analyzed through structured video coding by two authors and aligned with system logs on an event-by-event basis. 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.

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.

6 Study Results

6.1 Strategies for Integrating Multimodality in Creative Writing

To address RQ1, we identified four key interaction strategies employed when using Vistoria.

First, participants leveraged multimodal elements to externalize and interconnect different branches of their ideas. Second, they utilized image-text pairs as catalysts to generate richer and more dynamic narrative descriptions. Third, participants employed visual instruments as a mechanism to drive pivotal shifts in their storytelling. Finally, they treated images as both checkpoints and traces, enabling them to maintain narrative coherence and facilitate recall throughout the story development process.

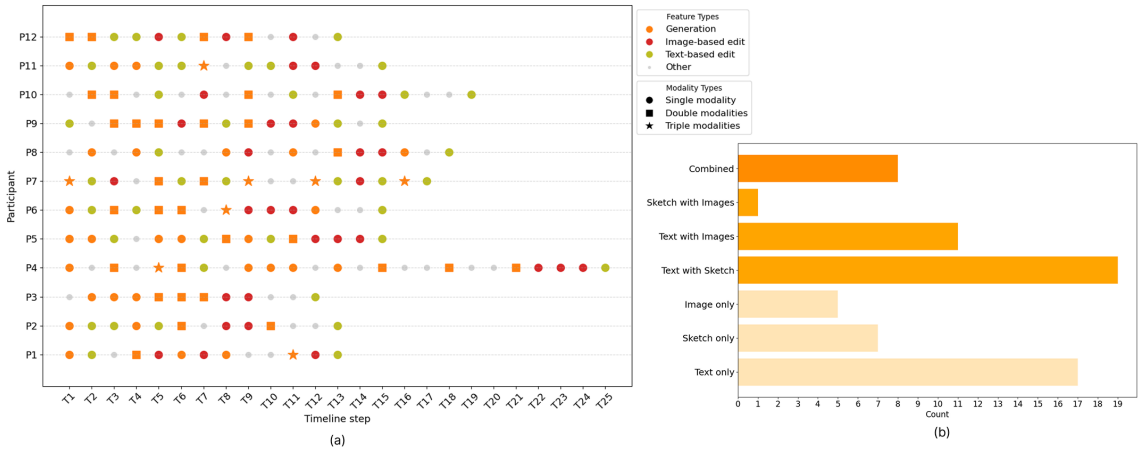

Fig. 6. Interaction records of all participants. The creative workflow begins with multimodal generation—primarily text, complemented by sketches or images to express intentions, followed by refinement and iteration using visual instruments, during which textual descriptions are continuously revised 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.

		Vistoria		Baseline		Statistics	
		mean	std	mean	std	p	Sig.
Transparent of ML model	Transparency	4.250	1.865	3.417	2.234	0.2015	–
	Controllability	4.750	1.357	4.833	1.586	0.8664	–
	Sense of Collaboration	5.583	1.505	4.333	1.670	0.0206	*
	Support for Thought Process	5.333	1.371	4.750	1.960	0.3172	–
	Match to goal	5.500	0.905	4.833	1.337	0.1201	–
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 2. Comparison of survey results: Vistoria vs. Baseline. Sig.: * $p < .05$; ** $p < .01$

6.1.1 Use Multimodality to Support Expressing and Connecting Ideas. Log data (Figure 6) and observations (Figure 13) reveal that participants followed a consistent creative workflow when using the system. The process typically began with multimodal content generation, where participants combined generation functions across multiple modalities to visualize their intended content. During this exploratory phase, participants iteratively collected and refined story elements through generated cards(images-text pairs), at the same time, frequently modifying text in the editor or regenerating images based on updated text.

As participants explored multiple visualization directions, they employed the Collage function to combine existing canvas content and discover new narrative possibilities, or used visual instruments (Lasso, Filter, Perspective Shifts) to simultaneously modify text and images. This process involved continuous comparison with LLM outputs to determine desired versus undesired elements (P1, P2, P4, P8, P10). Participants systematically highlighted useful images and text for final translation into the text editor, often using Cluster functions to summarize character or object settings for incorporation into their final writing.

Multimodal Expression Enhances Creative Intention. The multimodal integration enhanced participants' ability to convey creative intentions. As shown in Figure 6 (b), text served as the primary medium but was consistently supplemented with sketches and images for context, yielding significantly higher expressiveness ratings than the baseline ($M_{\text{Vistoria}} = 6.08$ vs. $M_{\text{Baseline}} = 4.33$, $p = .023$) (Table 2). 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."* (Figure 11). Nearly all participants (P2, P6, P10, P11) stressed that combining sketches with text and images aligned outputs more closely with their creative intentions. 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 is always different from what's in my head."* (Figure 7). 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 creative outputs.

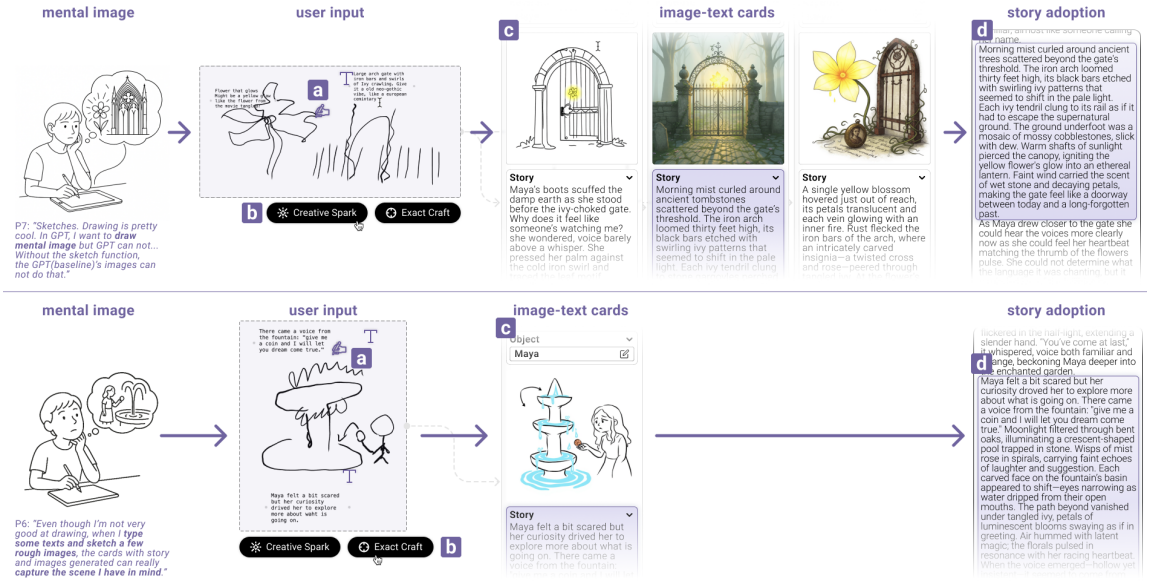

Fig. 7. Multimodal Expression. Example where sketch structure + textual details jointly yield multiple relevant images; participants valued sketches to visualize spatial layout beyond what text-only tools could provide.

Collage Function Enables Creative Recombination. The Collage function emerged as a central strategy for creative recombination, receiving positive usability ratings (M=5.33, SD=1.5) (Figure 8). 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 (Figure 9). 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."* 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.

Furthermore, multimodal input also increased participants' immersion with Vistoria compared to the baseline ($M_{\text{Vistoria}} = 4.92$ vs. $M_{\text{Baseline}} = 2.75$, $p = .0006$) (Table 2). This heightened immersion stemmed from the playful, enjoyable nature of multimodal creation and the emotional reinforcement of freely expressing intentions across modalities. As P1 reflected: *"Using the tool feels more like... doing something on a whiteboard or a big piece of paper, where*

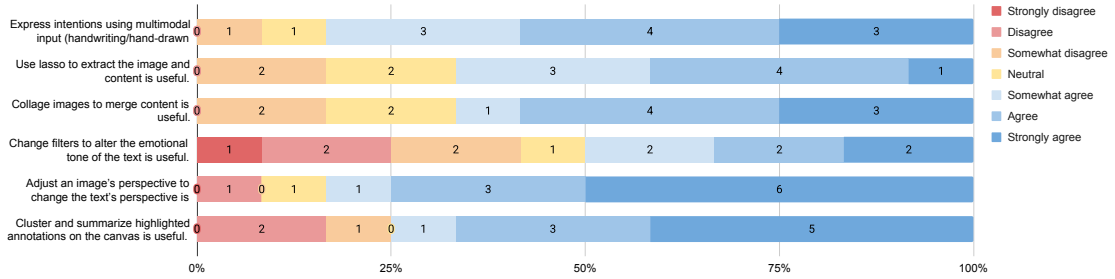

Fig. 8. The perceived usefulness of each key functionality in Vistoria

you can do almost anything—it's very free and interesting." Together, multimodal inputs and outputs fostered a more engaging and emotionally resonant creative experience.

6.1.2 Instrumental Interaction Tools Drive Narrative Development. Analysis of interaction logs (Figure 6 and 13) shows that participants often engaged with visual instruments in the mid-phase of creation, and these instruments directly shaped how stories were developed and revised.

Perspective Shift as a narrative frame-shifter. The *Perspective Shift* instrument, which altered both the image and the story perspective, was rated among the most useful features *(M=5.92, SD=1.56)* (Figure 8). 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. Changing the perspective of the voice of the story added personal sentiment and 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 10 (a)). Perspective Shift allowed authors to flexibly reconfigure narrative viewpoint and voice, surfacing new emotional framings and redirecting story trajectories without disrupting their ongoing writing flow.

Lasso as a granularity controller for local-to-global rewriting. The *Lasso* instrument was also valued as useful *(M=5.0, SD=1.28)* (Figure 8) for enabling writers 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 10 (b)) This reflects how the lasso tool enables a narrative "zoom" functionality, allowing writers to oscillate between macro-level story development and micro-level detail refinement within a single interface. Similarly, P9 and P11 described how the lasso allowed them to deepen or refine specific text and emphasize the 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.

Filters 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... After using the filter, I can't determine tone from text alone—I need to choose between suspenseful or romantic, and the visuals help me decide."* 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."* (Figure 10 (c)).

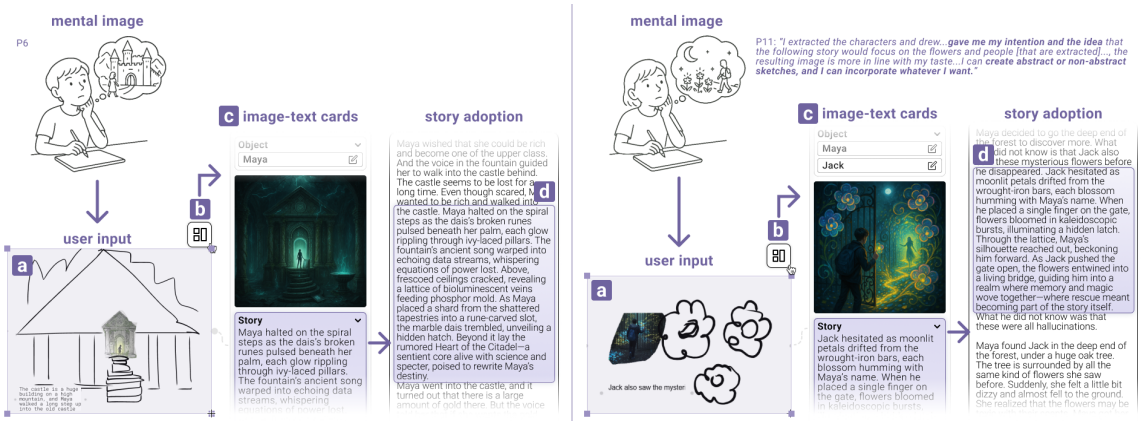

Fig. 9. Collage. Participants used Collage for creative recombination, merging extracted objects/scenes (often mixing sketches with images) to specify visualization and advance story ideas; usability was positively rated.

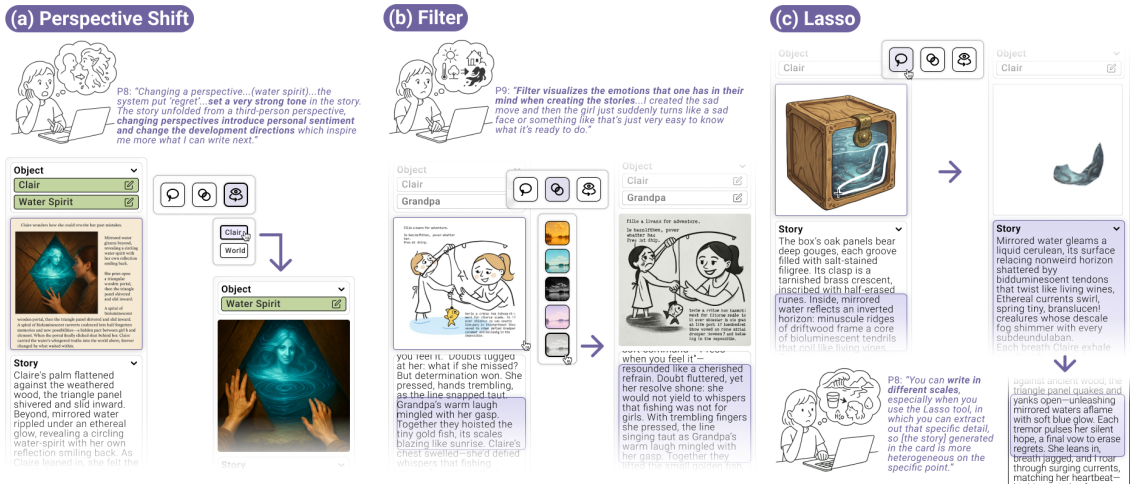

Fig. 10. (a) **Perspective Shift** was rated highly useful (M=5.92); changing the image viewpoint also reframed narrative voice and redirected story development, as participants reported; (b) **Filters** (M=4.25) synchronize mood and style across media—applying a visual filter also rewrites the associated text to match the intended emotional tone, ensuring stylistic coherence; (c) **Lasso** tool (M=5.0) enabled writers to focus at different narrative scales by extracting or isolating elements to steer local and global edits.

The immediate visual feedback fostered a more intuitive workflow, aligning emotional intent with visual representation. The immediate visual feedback aligned emotional intent with representation, streamlining tone-setting decisions.

Taken together, these tools acted as narrative instruments that transformed localized operations into meaningful shifts in viewpoint, scale, and tone. They enhanced authors' control, amplified expressiveness, and sustained a fluid perception–action loop throughout the creative process.

6.1.3 Structural Mapping between Images-Text Paris as a Catalyst for Vivid Story Writing. Participants emphasized that when text and images appeared together, images served as direct references during writing, easing detailed description

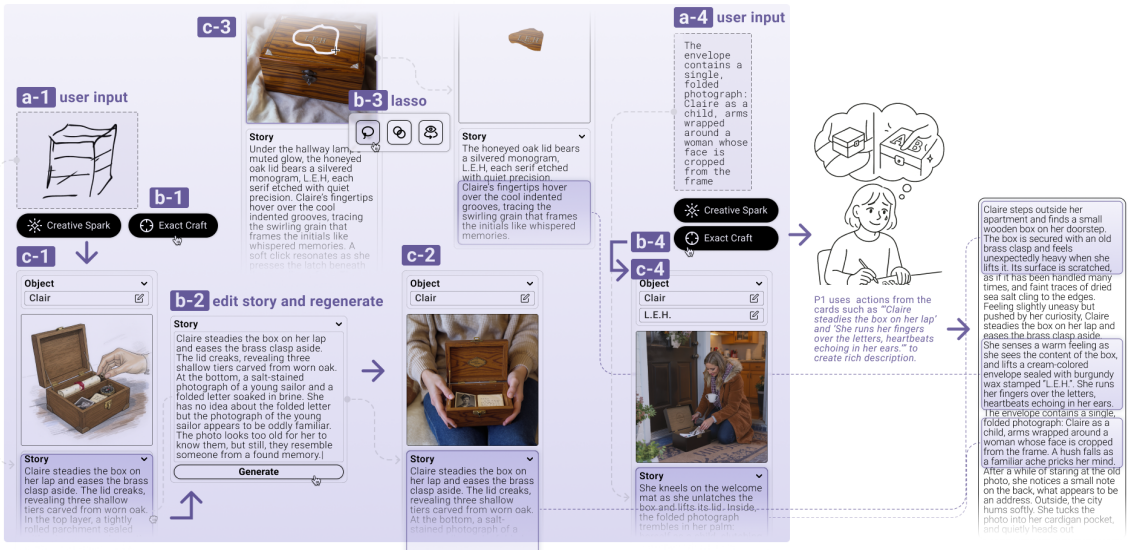

Fig. 11. Visual-to-text translation: writers turn visual cues into vivid prose that readers can picture; sample lines show how P1 used sensory-rich descriptions derived from an image-text pair.

and enriching narrative depictions. This integration worked through two cognitive mechanisms that supported vivid story construction.

Images-Text Paris as cognitive scaffolding for unfamiliar scenarios. Visual references helped participants imagine actions or settings beyond their lived experience. As P11 reflected, *"I'm more comfortable describing conversations and overall scenarios than people's actions—especially kids, who might have more interesting reactions. Visualizing with text descriptions helps me imagine unfamiliar actions, which I can then use in my story."* Here, the image acted as a structural anchor by aligning the visual cues of children's postures and expressions with possible textual descriptions. In other words, structural mapping between what was seen (visual structure) and what could be narrated (linguistic structure) scaffolded richer and more plausible action descriptions.

Images-Text Paris enabling descriptive language translation. Participants also converted visual elements into vivid prose, effectively mapping structures across modalities. As P5 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 final images in her writing and directly adopted text from the generated cards, integrating them into her final story. (Figure 12) This visual-to-textual translation was also evident in concrete writing outcomes. 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 *"She runs her fingers over the letters, heartbeats echoing in her ears."* to form his final story. (Figure 11) Here, the structural alignment between gesture in the image and embodied description in text allowed P1 to carry emotional undertones across modalities, turning static imagery into dynamic, sensory-rich narrative moments.

Overall, structural mapping between visual and textual representations scaffolded descriptive richness, enhanced vividness, and supported schema formation for scenarios participants might otherwise have struggled to imagine.

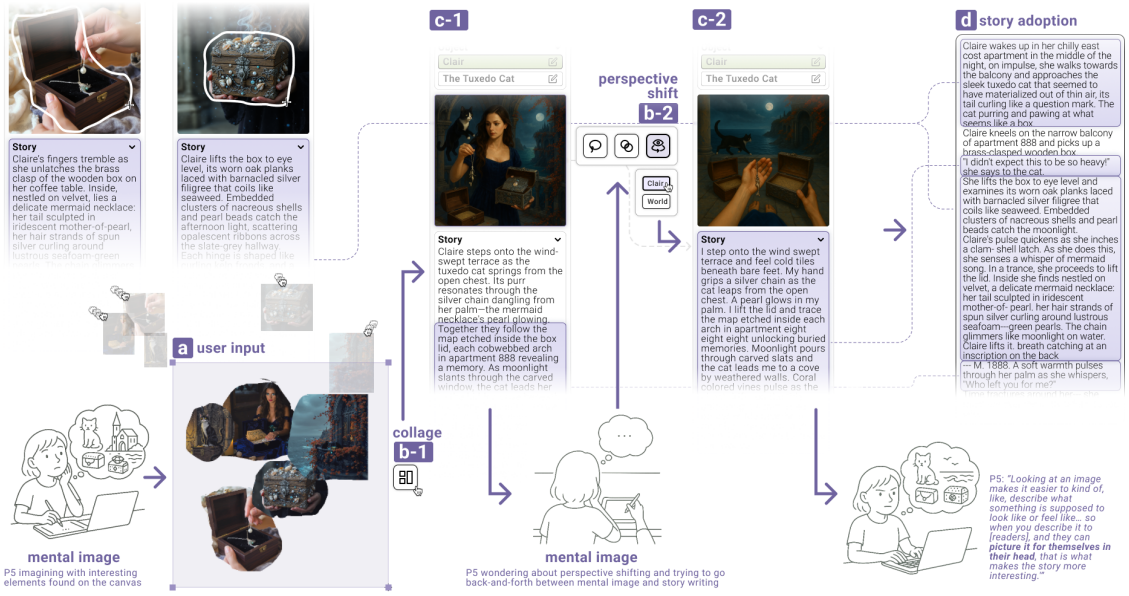

Fig. 12. Images act as cognitive scaffolds, helping writers describe unfamiliar actions or contexts more concretely. P5 constructed her final writing by referring to both images and adopted and edit related text.

6.1.4 Structural Mapping across Multiple Dimensions to Trace the Story. Participants used not only text but also images as visual checkpoints to trace story direction and development. Because images were more immediately perceptible than text, they helped participants align outputs with intended style or mood and, when viewed in sequence, detect progressions and gaps (P6, P7, P2). Within Vistoria, this process evolved into structural mapping: participants systematically compared visual–textual structures across dimensions, gaining multimodal visibility into narrative development.

This layered view created a relational space where main storylines and branching alternatives could be mapped against one another, allowing participants to perceive both continuity and divergence, thereby fostering coherence while supporting flexible redirection. As P12 noted, visuals alongside text set the mood and made it easier to recall earlier ideas, illustrating how structural mapping provided mnemonic scaffolding that supported continuity in storytelling. Compared to GPT's linear text-only workflow, participants found that Vistoria's visual traces better supported narrative management and reduced idea drift or loss (P12, P11, P2). P2 reflected on this benefit explicitly:

> *"Using the system, we started with a baseline picture and then came up with another picture… it was 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."*

Here, structural mapping across successive images enabled participants to align branches and selectively prune or redirect them. Such multi-dimensional mapping clarified narrative progression, improved recall and coherence, and supported flexible branching and gap detection, sustaining sensemaking throughout the evolving story.

	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 13, 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.

6.2 Multimodal Approaches Creation to Inspire Creativity

To answer RQ2, we found that image-text co-editing facilitated creativity in fictional story writing through three key cognitive processes: (1) divergent exploration, where participants spatially arranged multimodal components and pursued multiple narrative directions; (2) serendipitous discovery, where unexpected LLM outputs sparked new creative possibilities; and (3) multimodal cues activating memory and cultural connections. Together, these processes enhanced participants' sense of authorship by positioning the system as a collaborative partner that expanded creative possibilities rather than replacing human imagination.

6.2.1 Bottom-up Creation to Support Divergent Exploration. As shown in Figure 13 and Table 3, participants demonstrated significantly greater breadth and depth of exploration with Vistoria compared to the GPT baseline. They pursued a wider range of narrative directions ($M_{\text{System}} = 6.92, SD = 2.81$ vs. $M_{\text{Baseline}} = 1.42, SD = 1.08$) while also producing more variations within single paths (branches: $M_{\text{System}} = 3.00, SD = 1.35$ vs. $M_{\text{Baseline}} = 1.92, SD = 0.67$). Although the average depth per path was slightly lower for Vistoria ($M_{\text{System}} = 1.70, SD = 1.18$ vs. $M_{\text{Baseline}} = 2.00, SD = 1.22$), this reflected participants' substantially broader exploration overall; within individual directions, they often engaged in deeper iterations, as evident in Figure 13. Such enhanced exploration transformed their creative process from passively consuming LLM outputs to actively curating multimodal components.

The Vistoria canvas functioned as an exploratory space where participants pursued multiple storylines in parallel without linear constraints. This freeform environment encouraged spatial arrangement, clustering, and repositioning of ideas through manipulation and recombination of alternative story–image cards to explore divergent scenarios (P7, P4, P5, P8). 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. This exploratory approach helped participants avoid early fixation and sustained creative engagement. As P8 observed: *"GPT workflow is more streamlined... top-down. Using the visual app feels more bottom-up. You are open to possibilities, so there's not a finite result and more possibilities being explored."* Ultimately, this shift highlights how multimodal, spatially organized exploration fosters divergent thinking, sustains creative momentum, and supports a more generative form of narrative authorship.

6.2.2 Serendipity from LLM Randomness Enriches Narrative Development. Another recurring theme was that the inherent randomness of LLM-generated image-text output often introduced serendipitous elements into participants' creative process. Rather than being perceived as noise, these unexpected details were frequently embraced as inspiration, enriching story development and occasionally opening entirely new narrative directions.

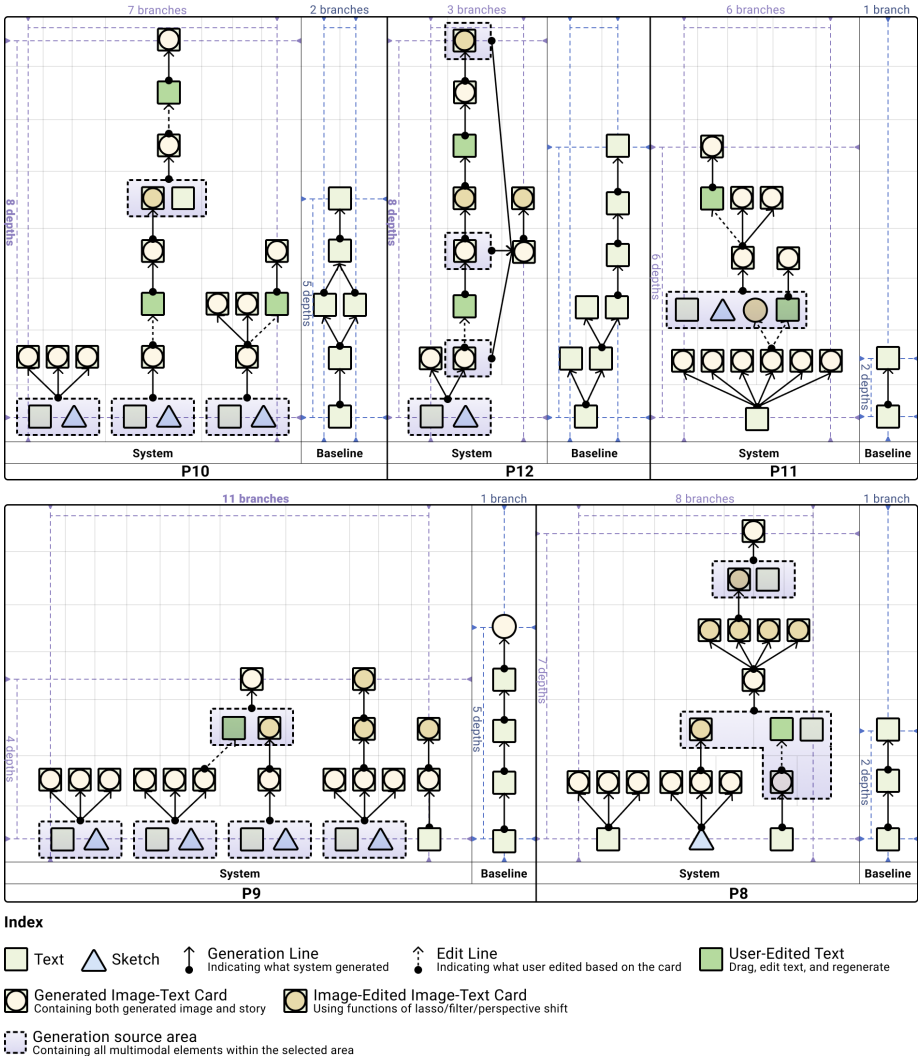

Fig. 13. Behavioral diagram contrasting Vistoria vs. baseline: participants cycle through multimodal generation, collage/recombination, and coupled image–text editing before collecting highlights for integration. Data from P8-P12 show that, compared with baseline, participants using Vistoria explored more directions, with greater divergence (Branches) within each direction. Participants also tended to pursue deeper exploration within specific directions when using Vistoria.

For example, when P1 generated an image-text pair of the mysterious box, the system unexpectedly inserted the logo "LEH." on the box and in the text description. Intrigued, P1 subsequently asked the system to visualize who or what "LEH" could be, gradually evolving this accidental motif into a recurring symbol within their final story (Figure 11). Nearly all participants described how surprises in LLM-generated images or texts gave them fresh possibilities to explore. P12 captured this dynamic vividly: *"The system gave me a riddle at an early stage, I didn't understand it at first, but as I kept working, I slowly decoded it, and this process of unraveling became part of my story. It helped me construct something unique that I wouldn't have created otherwise."* (Figure 14) Together, these examples show how the interplay

of multimodal generation and LLM's inherent unpredictability fueled creativity. Randomness supported divergence by surfacing novel ideas, while participants retained agency in selectively integrating these elements, leading to unique, personally authored narratives.

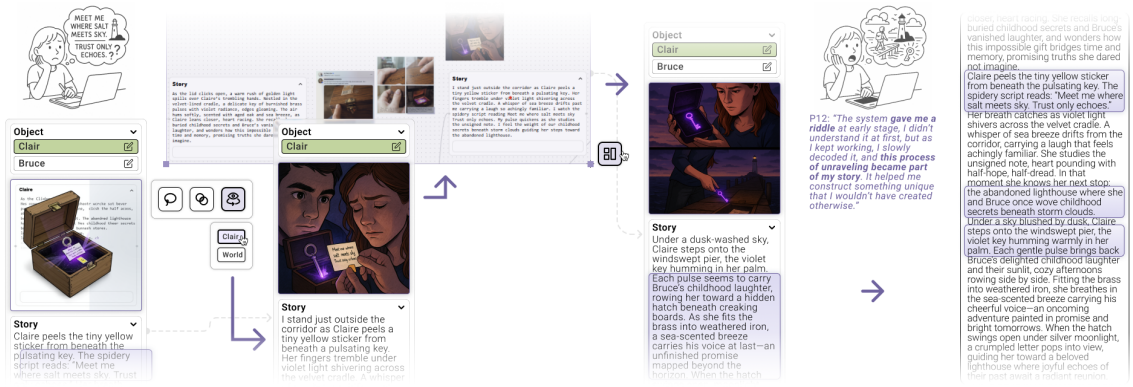

Fig. 14. In the creation process of P12, unexpected LLM randomness served as serendipitous prompts: a generated riddle was decoded and transformed into a key-and-note motif guiding Claire toward the lighthouse, illustrating how stochastic details became generative cues that enriched development and redirected the plot.

6.2.3 Unlocking Creativity Through Memory and Cultural Resonance. Participants reported that multimodal cues routinely triggered personal memories and cultural references, enriching their stories. Visual props, e.g., a medieval mask, readily invoked film/literary associations; as P4 noted, *"Seeing the mask immediately brought Batman/Phantom to mind... it made choosing a style much easier."* (Figure 15) Crucially, image–text pairs co-triggered lived experiences. In P7's case, a fountain scene image together with the phrase "mossy stone" jointly evoked a vivid memory, which she then translated into a concrete narrative element, the envisioned fall that became her story's ending(Figure 15). These multimodal traces turned associative recall into usable story material, weaving authenticity and imaginative depth into the narrative.

6.2.4 Strengthened Authorship and Agency. Enhanced exploration with Vistoria strengthened participants' sense of authorship. While GPT outputs in the baseline condition were consistently described as "surface-level" (P2, P5, P7) and felt like "someone else's work" (P3), almost all participants reported markedly stronger authorship with Vistoria, enabling deeper exploration and richer plots (P1, P2, P3, P5, P12, P10). P5 articulated this shift:

> *"When using Vistoria, every idea originated from my own imagination, and the final story was formed by combining these different self-generated ideas. This gave me a strong sense that the story was truly my own creation. I think of the system less as a generator and more as a companion sketchbook. While you explore different directions, you always remain in charge of the final piece."*

P9 reinforced this sentiment: *"This tool is more 'me'... I control characters and plots; ChatGPT may not become my story."* Quantitative results corroborated these perceptions: participants rated Vistoria as providing stronger collaborative experiences (CSI: $M_{\text{Vistoria}} = 5.50$ vs. $M_{\text{Baseline}} = 4.58$, $p = .0418$) (Table 2). Participants framed Vistoria as a supportive co-pilot rather than a replacement; as P10 noted: *"I consider this system as a copilot that does not help you with everything,*

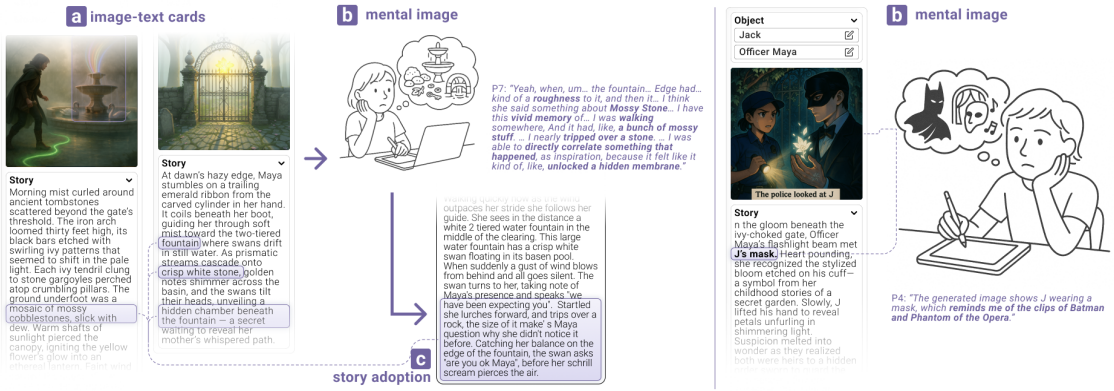

Fig. 15. Evolving cards leave visual traces that cue personal memories and cultural references (e.g., a mask → Batman/Phantom; a "mossy stone" → place memory), which writers externalize into concrete scenes. These multimodal traces act as mnemonic scaffolds, helping retrieve earlier thinking and weave lived experience into the story world—transforming associative recall into narrative material.

and this system that you have to create by yourself." Together, these findings mark a transition from passively adopting others' ideas to actively curating one's own generative outputs, reinforcing creative agency and ownership.

6.3 Tradeoffs of Multimodal Interaction for Creative Writing

While image–text co-editing enhanced authorship, coherence, and expressiveness, it also increased workload. NASA–TLX scores showed significantly higher mental demand ($M_{\text{Vistoria}} = 5.16$ vs. $M_{\text{baseline}} = 3.17$, p=.0000) and physical demand ($M_{\text{Vistoria}} = 4.67$ vs. $M_{\text{baseline}} = 2.08$, p=.0002), plus moderately higher frustration ($M_{\text{Vistoria}} = 2.75$ vs. $M_{\text{baseline}} = 1.75$, p=.0204) (Table 2). These increases reflect the effort of coordinating across modalities and actively curating outputs relative to the GPT baseline. 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): *"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 arose from unfamiliarity with the canvas interface (P3, P10).

6.3.1 Higher burden of authorship. Almost all participants valued the ability to maintain control of the story, and the creation process gave them a stronger sense of ownership. Yet, this empowerment came with a cost: several participants 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 required users to supply and elaborate on their own ideas before meaningful generation occurred, which may lead 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 receiving GPT's ready-made text (P1, P2, P11). Thus, stronger authorship agency came at the cost of higher workload.

6.3.2 Validating ideas rather than generating them. Some participants noted that the system's strengths lay in validating or expanding concrete mental images rather than generating 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 offering prompt-like guidance, whereas the multimodal Vistoria system primarily served to elaborate or diverge from existing visions. This made the multimodal approach particularly valuable for writers with partially formed concepts, but less helpful during the most open-ended phases of ideation when abstract exploration is more important than visual specificity.

6.3.3 Cognitive offloading and efficiency despite higher load. Despite a higher mental workload, participants repeatedly emphasized that multimodality enabled more efficient allocation of attention. From participants' view, the image-text pairs externalized fleeting ideas, preserved spatial and sensory detail, and reduced information loss when translating imagination into tangible artifacts (P4, P7). As P7 noted, *"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."*

Thus, while Vistoria required more effort and decision-making, it also freed participants from lower-level memory maintenance and allowed them to focus on higher-order creative synthesis. With greater familiarity, the efficiency gains from this offloading may outweigh the initial overhead.

7 Discussion

7.1 Cross-modal LLM-Mediated Instrumental Interaction

We extend the notion of Instrumental Interaction into an LLM-mediated, cross-modal paradigm where generated artifacts repeatedly re-enter the workflow as manipulable outputs. In classic formulations, instruments mediate between users and domain objects, maintaining a separation between "tool" and "content" [9]. In generative contexts, this dichotomy breaks down: each card (image–text pair) is simultaneously an outcome and a renewed object of operation, enabling recursive, non-linear composition [59].

This reframing also reshapes reification. Vague creative intents expressed through text, sketches, or images materialize across modalities into manipulable, cross-linked artifacts that remain editable and reusable. Reification thus spans modalities rather than existing in a single channel, supporting direct manipulation and reflection on both structure and semantics of evolving narratives [46, 59].

A key mechanism enabling this circulation is structural mapping [27, 46, 52]. Participants made sense of their evolving materials not only by inspecting the surface properties of each medium but also by exploiting the correspondences in shared properties, concepts, or identities established between them. Refining a visual scene, for example, carried implications for narrative structure, tone, or pacing in text, while textual edits translated into compositional adjustments in images. In this way, mappings transformed what is usually a cognitive act of alignment into an operational resource: they stabilized coherence while keeping alternatives visible and accessible. Instrumental interaction and structural mapping thus operate as mutually reinforcing processes. Instruments generate the artifacts that mappings align, while mappings scaffold the interpretability and reuse of instruments.

This dynamic carries several design implications for future systems. They should not only offer isolated multimodal tools but cultivate instrumental ecosystems in which mappings are first-class and inspectable, and where families of instruments carry coherent semantics across domains [59]. Furthermore, instead of overwhelming users with immediate complexity, operations can be gradually disclosed, allowing mappings to surface progressively and adapt to

context [2, 65]. Provenance views that expose the cascade of edits across modalities could support both divergence and recombination, helping individuals and teams understand how narratives branch and converge [46].

Ultimately, the LLM can be positioned less as a generator of content and more as a mediator of instruments that align structures, propagate semantics, and synthesize new operations from prior materials. This shift reimagines authorship as an orchestration of evolving, cross-modal instruments, opening a design space where exploration and coherence can be sustained simultaneously [46, 62].

7.2 Orchestration and Curatorial Authorship to Form Creativity in Cross-Modal Creation

Traditional creativity theories emphasize individual expression through language and pre-formed intent [21, 66], positioning the creator as the primary originator of content. Our findings challenge this view by demonstrating that creativity in *Vistoria* manifested as orchestration. Participants conducted and harmonized streams of multimodal LLM outputs through selection, combination, and reframing. Creative goals emerged through interaction with generated materials rather than preceding them, with meaning composed across representational boundaries. This suggests creativity in AI-augmented contexts should be theorized as the strategic coordination of evolving possibilities, repositioning authorship from origination to curation and synthesis within a collaborative ecosystem.

While existing multimodal creativity research often focuses on seamless integration between modes [60, 63, 80], our study reveals that productive tensions between text and image became central to the creative process. Participants actively negotiated the trade-offs between imagistic specificity and narrative openness, treating these tensions not as problems to solve but as resources for exploration. The ongoing adjudication process—accept, adapt, defer, or discard–became the primary engine of creative coherence. This extends theories of creative constraint by showing how representational tensions across modalities can drive rather than hinder creative development.

Furthermore, our findings contribute to debates about human agency in LLM-assisted creation by demonstrating that curatorial work enhanced rather than diminished participants' sense of ownership. Authorship was experienced as steering a dialogue with the model, including deciding what counts, where elements belong, and how disparate fragments accrue into a unified voice. This sense of agency grew precisely from having meaningful choices to negotiate among viable alternatives [57]. This suggests a new model of distributed authorship where human control emerges through deliberate curation within a space of LLM-generated possibilities.

As for design implications, our findings suggest that future system design should prioritize support for orchestration over seamless, one-shot generation. Instead of producing singular outputs, tools could by default surface multiple cross-modal alternatives, accompanied by lightweight mechanisms that enable users to accept, adapt, defer, or discard options as part of an ongoing multimodal curatorial process [57]. Moreover, systems should not only present outputs but also make the inherent tensions between modalities visible and adjustable. This can be done, for example, through interactive controls such as ambiguity sliders that allow users to navigate between imagistic specificity and narrative openness.

Beyond simple interface controls, future systems could also provide deeper support for curatorial authorship. For example, they might keep track of where each fragment or idea came from across different iterations, or offer reusable "curation moves" that let users quickly repeat and adapt common ways of combining or editing materials [64]. In this way, design emphasis shifts away from achieving frictionless integration and toward helping users actively steer evolving creative possibilities, foregrounding authorship in AI-mediated contexts as a matter of curation, synthesis, and negotiated choice [20].

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 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 the 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 on conducting field studies to 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. 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 participant pool 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 professional authors, older and younger writers, and individuals from varied cultural backgrounds, to more fully evaluate the system's applicability and generalizability.

8 Conclusion

This paper introduced Vistoria, a multimodal text–image co-editing system designed to support the cyclical "planning–translation" process in fictional writing. Grounded in Wizard-of-Oz co-design study and theories of Instrumental Interaction and Structural Mapping, the system integrates four key capabilities: (1) representing sketches, texts, and references as manipulable card units; (2) enabling cross-modal alignment so that edits in one modality propagate to the other; (3) providing polymorphic operations that apply consistently across text and image; and (4) supporting organization and reuse through highlighting and clustering panels. A controlled study with 12 creative writers demonstrated that Vistoria enhanced writers' expressiveness, immersion, and sense of collaboration, while broadening exploration of narrative branches and reinforcing authorship and control. These findings suggest that Vistoria transforms the writing process into a dialogic interaction with multimodal instruments, enabling authors to negotiate meaning and control across modalities.

References

[1] Safinah Ali and Devi Parikh. 2021. Telling creative stories using generative visual aids. *arXiv preprint arXiv:2110.14810* (2021).

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

[3] 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.

[4] Anthropic. 2025. Claude.ai — New Chat. https://claude.ai/new. Accessed: 2025-08-29; may require login.

[5] Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L Glassman. 2024. Chainforge: A visual toolkit for prompt engineering and llm hypothesis testing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

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

[7] 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.

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

[9] 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

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

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

[12] 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.

[13] 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

[14] 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 *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–19. https://doi.org/10.1145/3491102.3501819

[15] 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*. 1–19.

[16] 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

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

[18] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies: why and how. In *Proceedings of the 1st international conference on Intelligent user interfaces*. 193–200.

[19] 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

[20] Zijian Ding. 2024. Advancing gui for generative ai: Charting the design space of human-ai interactions through task creativity and complexity. In *Companion Proceedings of the 29th International Conference on Intelligent User Interfaces*. 140–143.

[21] Dianne Donnelly. 2011. *Establishing creative writing studies as an academic discipline*. Vol. 7. Multilingual Matters.

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

[23] 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.

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

[25] Kexue Fu, Ruishan Wu, Yuying Tang, Yixin Chen, Bowen Liu, and Ray Lc. 2024. " Being Eroded, Piece by Piece": Enhancing Engagement and Storytelling in Cultural Heritage Dissemination by Exhibiting GenAI Co-Creation Artifacts. In *Proceedings of the 2024 ACM designing interactive systems conference*. 2833–2850.

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

[27] Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science* 7, 2 (1983), 155–170.

[28] 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

[29] Katy Ilonka Gero, Chelse Swoopes, Ziwei Gu, Jonathan K Kummerfeld, and Elena L Glassman. 2024. Supporting sensemaking of large language model outputs at scale. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–21.

[30] Elena L Glassman, Jeremy Scott, Rishabh Singh, Philip J Guo, and Robert C Miller. 2015. OverCode: Visualizing variation in student solutions to programming problems at scale. *ACM Transactions on Computer-Human Interaction (TOCHI)* 22, 2 (2015), 1–35.

[31] Elena L Glassman, Janet Sung, Katherine Qian, Yuri Vishnevsky, and Amy Zhang. 2018. Triangulating the news: Visualizing commonality and variation across many news stories on the same event. (2018).

[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] 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.

[35] Mahmood Jasim, Christopher Collins, Ali Sarvghad, and Narges Mahyar. 2022. Supporting serendipitous discovery and balanced analysis of online product reviews with interaction-driven metrics and bias-mitigating suggestions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–24.

[36] 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.

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

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

[39] Max Kreminski, John Joon Young Chung, and Melanie Dickinson. 2024. Intent Elicitation in Mixed-Initiative Co-Creativity.. In *IUI Workshops*.

[40] 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, et al. 2024. A design space for intelligent and interactive writing assistants. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–35.

[41] 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*. 1–18.

[42] 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.

[43] LING Long, CHEN Xinyi, WEN Ruoyu, LI Toby Jia-Jun, and LC Ray. 2024. Sketchar: supporting character design and illustration prototyping using generative AI. *Proceedings of the ACM on Human-Computer Interaction* 8, CHI PLAY (2024), 337.

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

[45] Damien Masson, Young-Ho Kim, and Fanny Chevalier. 2024. Textoshop: Interactions Inspired by Drawing Software to Facilitate Text Editing. https://doi.org/10.48550/arXiv.2409.17088 arXiv:2409.17088 [cs]

[46] Damien Masson, Zixin Zhao, and Fanny Chevalier. 2024. Visual Writing: Writing by Manipulating Visual Representations of Stories. *arXiv preprint arXiv:2410.07486* (2024). https://doi.org/10.48550/ARXIV.2410.07486

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

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

[49] Inc. Midjourney. 2025. *Midjourney — Home.* https://www.midjourney.com/home Homepage of the independent research lab Midjourney that explores new mediums of thought.

[50] 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

[51] 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.

[52] Don Norman. 2013. *The design of everyday things: Revised and expanded edition*. Basic books.

[53] OpenAI. 2024. *GPT-4o*. https://openai.com/index/hello-gpt-4o/

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

[55] 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*. 1–19.

[56] 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

[57] Mitchel Resnick, Brad Myers, Kumiyo Nakakoji, Ben Shneiderman, Randy Pausch, Ted Selker, and Mike Eisenberg. 2005. Design principles for tools to support creative thinking. (2005).

[58] 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).

[59] 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

[60] 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

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

[62] Keith Sawyer. 2017. *Group genius: The creative power of collaboration*. Basic books.

[63] 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*. 1–20.

[64] Ben Shneiderman. 2007. Creativity support tools: accelerating discovery and innovation. *Commun. ACM* 50, 12 (2007), 20–32.

[65] Ben Shneiderman and Pattie Maes. 1997. Direct manipulation vs. interface agents. *interactions* 4, 6 (1997), 42–61.

[66] Hazel Smith. 2020. *The writing experiment: strategies for innovative creative writing*. Routledge.

[67] 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*. 1–26.

[68] JD Swerzenski. 2021. Fact, fiction or Photoshop: Building awareness of visual manipulation through image editing software. *Journal of Visual Literacy* 40, 2 (2021), 104–124.

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

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

[71] 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*. 1–28.

[72] 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*. 1–26.

[73] 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

[74] 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).

[75] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–22.

[76] Tongshuang Wu, Kanit Wongsuphasawat, Donghao Ren, Kayur Patel, and Chris DuBois. 2020. Tempura: Query analysis with structural templates. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[77] 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*. 853–866.

[78] Litao Yan, Miryung Kim, Bjoern Hartmann, Tianyi Zhang, and Elena L Glassman. 2022. Concept-annotated examples for library comparison. In *Proceedings of the 35th annual ACM symposium on user interface software and technology*. 1–16.

[79] 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

[80] 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

[81] Ryan Yen, Jiawen Zhu, Sangho Suh, Haijun Xia, and Jian Zhao. 2023. Coladder: Supporting programmers with hierarchical code generation in multi-level abstraction. *arXiv preprint arXiv:2310.08699* (2023).

[82] 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 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.2 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	Emphasizes 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.3 LLM Prompts

Communication between different agents is carried out through JSON-formatted messages, ensuring efficient and reliable information transfer.

All prompts are structured in the following format:

- Blue text indicates placeholders that will be replaced with task-specific information in different generation requests.
- Pink text represents the mandatory output type of information generated by the agent.

The below is an example prompt for precise description generation of GPT-4o.

```
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.
```