Spatial Balancing: Harnessing Spatial Reasoning to Balance Scientific Exposition and Narrative Engagement in LLM-assisted Science Communication Writing

ANONYMOUS AUTHOR(S)

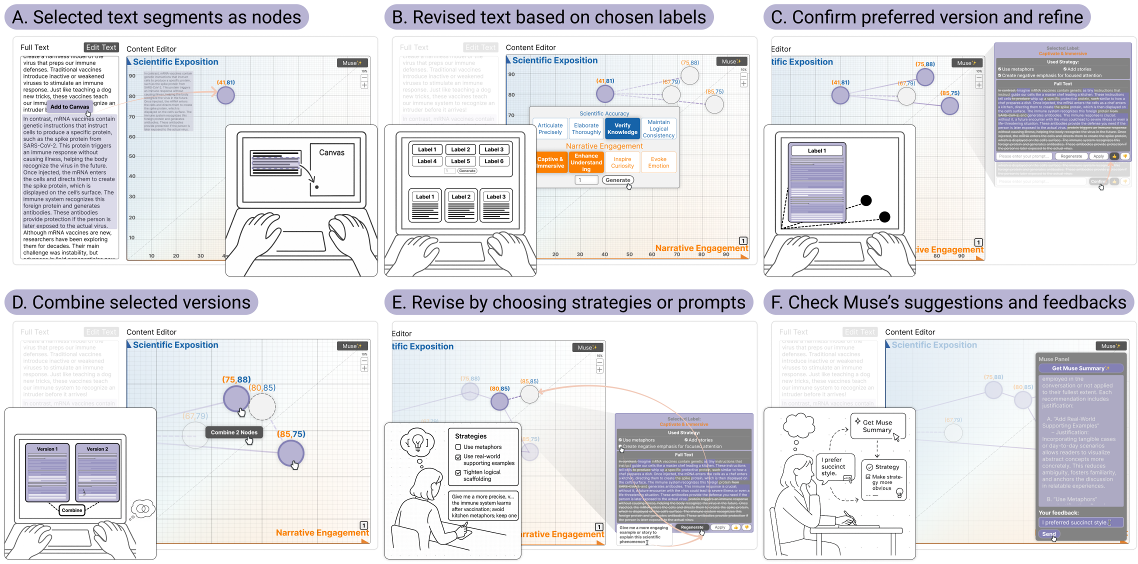

Fig. 1. Example Workflow of using SpatialBalancing for iterative science communication writing. A – Jenny drags her draft into the canvas, where each paragraph becomes a node mapped by Scientific Exposition (Y-axis) and Narrative Engagement (X-axis). B – She selects revision labels such as Enhance Understanding or Captivate & Immerse, each tied to LLM-driven strategies that generate new versions placed accordingly. C – Jenny reviews and confirms preferred revisions, which turn purple for further refinement. D – She can combine two versions into a synthesized draft, balancing credibility and engagement. E – Further revisions are guided by strategies or custom prompts, enabling precise, iterative control. F – Finally, SpatialBalancing's Muse assistant reflects on her revision history and offers adaptive suggestions.

Balancing scientific exposition and narrative engagement is a central challenge in science communication. To examine how to achieve balance, we conducted a formative study with four science communicators and a literature review of science communication practices, focusing on their workflows and strategies. These insights revealed how creators iteratively shift between exposition and engagement but often lack structured support. Building on this, we developed SpatialBalancing, a co-writing system that connects human spatial reasoning with the linguistic intelligence of large language models. The system visualizes revision trade-offs in a dual-axis space, where users select strategy-based labels to generate, compare, and refine versions during the revision process. This spatial externalization transforms revision into spatial navigation, enabling intentional iterations that balance scientific rigor with narrative appeal. In a within-subjects study (N=16), SpatialBalancing enhanced metacognitive reflection, flexibility, and creative exploration, demonstrating how coupling spatial reasoning with linguistic generation fosters monitoring in iterative science communication writing.

Manuscript submitted to ACM

CCS Concepts: • **Human-centered computing → Collaborative interaction**.

Additional Key Words and Phrases: Narrative Strategy, Science Communication, Spatial Reasoning, Writing Assistance

ACM Reference Format:
Anonymous Author(s). 2026. Spatial Balancing: Harnessing Spatial Reasoning to Balance Scientific Exposition and Narrative Engagement in LLM-assisted Science Communication Writing. In *CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 36 pages. https://doi.org/10.1145/3474349.3480203

1 Introduction

With the recent progress in modeling human language, generative systems have increasingly been applied to diverse writing tasks, ranging from news reporting to creative storytelling [54]. Compared with other forms of writing, science communication is distinctive in that it requires a careful balance between scientific exposition and narrative engagement, a balance that directly influences how the public understands and trusts scientific knowledge [7, 16, 29, 78].

Online platforms have democratized science content creation across YouTube, social media, blogs, Q&A sites, and podcasts [67, 92, 99], making the balance between scientific exposition and narrative engagement increasingly complex. While improving accessibility, this trend creates a significant challenge: untrained creators produce highly variable content quality [76], highlighting the need for better guidance frameworks to support high-quality science communication. To address this challenge, recent HCI research has leveraged the linguistic intelligent of large language models (LLMs) to support science communication writing, capitalizing on their ability to synthesize complex information, switch flexibly between tones, and produce stylistic alternatives [11]. These systems focus on content planning [72, 80], rhetorical enhancement [34, 35, 49], and iterative revision [58, 98]. However, existing tools predominantly adopt a prompt–response paradigm, offering surface-level language variation while focusing on either structural planning [80] or localized iterations [36, 49]. They lack integration between local edits and broader narrative design, and provide no structured representations showing how revisions influence the trade-off between exposition and engagement, constraining users' capacity to intentionally guide this process [98].

These limitations point to a broader question: as generative systems increasingly approximate or surpass human linguistic capabilities, effective human–AI collaboration may need to draw on distinctively human cognitive capacities, such as spatial reasoning. Spatial reasoning, defined as the ability to comprehend, manipulate, infer, and anticipate spatial relationships and structures [56], enables holistic navigation of complex configurations. HCI research has leveraged spatial representations for human–AI collaboration through node-link diagrams [49, 61, 100] and selection paradigms [58, 59] to enhance interpretability and controllability. However, these approaches remain limited to structural visualization and option manipulation, lacking mechanisms for dynamically sustaining the balance between scientific exposition and narrative engagement—an inherently ongoing, multidimensional trade-off [25, 38, 62]. This balancing process resembles spatial navigation, where writers must continually evaluate their position relative to communicative goals and make directional adjustments [95, 98], underscoring the need for spatial-reasoning-based interfaces that better support communicators in navigating this balance.

Building on this foundational spatial reasoning capacity of human and LLM's linguistic intelligence, we propose **Spatial Balancing**: an interaction paradigm that leverages human spatial reasoning to guide the dynamic negotiation between scientific exposition and narrative engagement in science communication. Communicators use spatial reasoning to steer and regulate the dynamic negotiation between scientific exposition and narrative engagement, while LLMs provide the linguistic material that fills this rhetorical space [14]. We instantiate this concept in SpatialBalancing, a proof-of-concept system (Figure 1) that visualizes communicative iterations with LLM in a two-dimensional coordinate space, where the x-axis represents narrative engagement and the y-axis represents scientific exposition. Each iteration is plotted as a point, thereby providing communicators with continuous visual feedback to assess how revisions shift across the two dimensions. Such spatial externalization reframes revision as a process of navigating a rhetorical space, shifting the activity from reactive modification toward more intentional exploration. This idea further aligns with recent work on LLM-assisted ideation that employs spatial representations to scaffold scientific ideation [23].

To systematically investigate this approach, we pose three research questions:

- **RQ1 - System Design:** How can spatial reasoning be applied to support writers in balancing scientific exposition and narrative engagement in science communication?
- **RQ2 - Cognition:** What impact does spatial visualization of revision tradeoffs have on writers' cognitive process?
- **RQ3 - System Feature:** How do different interface features (2D coordination, strategy labels, reflective feedback) contribute to improving writing quality and user experience in science communication?

In a within-subjects study with 16 science communicators, SpatialBalancing demonstrated measurable advantages over a strong LLM baseline. It supported greater strategic flexibility, creative exploration, and metacognitive reflection. Participants reported that the coordinate visualization externalized abstract goals, facilitated real-time self-monitoring, and enhanced their confidence in editorial decisions. By making the trade-offs between scientific exposition and narrative engagement more tangible, the system enabled more deliberate decision-making and iterative exploration.

Our contributions include:

- The concept of Spatial Balancing as a novel interaction paradigm for science communication writing, along with design implications that translate spatial reasoning into actionable writing support with LLM.
- A proof-of-concept system that instantiates this framework through spatial reasoning, enabling visual exploration of revision trade-offs.
- Empirical evidence from a within-subjects study with 16 science communicators showing that our proof-of-concept system improves metacognitive regulation, creative exploration, and writer confidence relative to a state-of-the-art LLM baseline.

2 Related Work

2.1 Balancing Exposition and Narrative Engagement in Science Communication Writing

In the Information Age, online science communication has become increasingly dominant, especially in the popular science field [9, 63]. Science communication refers to the strategic use of various forms of communication, such as media, events, and interactions, to convey scientific information to diverse audiences in a way that aims to increase awareness, enjoyment, interest, opinion-forming, and understanding [7, 46, 66]. The popular science movement (also known as pop science or popsci) aims to interpret and present scientific concepts in an accessible way for a general audience, placing greater emphasis on entertainment and broadening its scope compared to traditional science journalism [5, 19, 93]. As online communication technologies have become more accessible, various formats have emerged to deliver popular science content, including books, documentaries, web articles, and online videos [29, 93, 99].

A fundamental challenge in science communication writing lies in balancing two often competing dimensions: scientific exposition and narrative engagement [25, 38, 62]. Burns et al. [7] made a vivid analogy, describing science communication writing as a form of "mountain climbing," balancing between scientific literacy and science culture.

Similarly, Dahlstrom [16] emphasized that science communication writing inherently involves both narrative and expository elements. In this study, we use the terms "scientific exposition" and "narrative engagement" to describe this tradeoff [24], because these terms more directly capture the practical tension between maintaining rigorous, detailed scientific presentation and creating compelling, accessible content for diverse audiences [24, 62]. The tension between these dimensions stems from their fundamentally different linguistic requirements. Engaging content relies on narrative techniques—storytelling, analogy, and suspense—to capture attention [16, 29, 38], while scientifically accurate content demands rigorous expository writing that prioritizes scientific detail and credibility [48, 51].

To address this inherent tension, writers typically navigate between these two dimensions using iterative linguistic strategies [29, 33, 64, 69], transforming revision into a non-linear, multi-pass process. Existing scholarship has developed strategies that focus on either narrative engagement or scientific precision [3, 52]. For enhancing narrative engagement, research has identified three primary approaches. First, writers create memorable points by distilling complex ideas into condensed, succinct expressions [29, 64]. Second, they evoke emotions by strategically incorporating elements of hope, fear, or sadness [32, 33, 42, 91]. Third, they spark curiosity through thought-provoking questions that encourage reader reflection [77, 99]. In contrast, strategies for maintaining scientific precision emphasize rigorous expository writing that prioritizes comprehensive detail and establishes credibility [48, 51, 69]. Through iteratively revising and evaluating drafts, writers achieve overall balance by strategically emphasizing engagement in some sections while prioritizing scientific exposition in others to ensure clear explanation throughout the piece [3, 43].

Most critically, science communication authors revise without timely, reader-centered feedback on how their text balances exposition and engagement [95, 98, 99]. This evaluation gap obscures whether a change represents an improvement or regression, pushing writers toward conservative edits and stifling exploration [3, 65]. Without reliable, localized signals, they must navigate implicit trade-offs that remain difficult to surface and track, creating subsequent challenges in judging whether revisions enhance the balance and generating reluctance to pursue alternatives due to fear of losing progress [98].

Existing approaches exacerbate the problem. Theory-heavy guidance provides minimal procedural support for iterative revision that balances scientific exposition with narrative engagement [25, 38, 62]. Consequently, there is a critical need for integrated, revision-oriented support that makes both dimensions visible across multiple scales, delivers real-time audience-informed feedback, and enables multi-version exploration through non-linear history with granular controls.

2.2 Spatial Reasoning for Steering Linguistic Intelligence of Language Models

Human cognition embodies two complementary strengths: linguistic intelligence, the capacity to generate, interpret, and manipulate complex symbolic expressions, and spatial reasoning, which supports envisioning relationships, operating on conceptual structures, and weighing trade-offs across multiple goals in multi-dimensional space. With recent advances, large language models (LLMs) have demonstrated remarkable linguistic intelligence—synthesizing complex information, flexibly shifting between tones, and producing stylistic alternatives that rival or even surpass human fluency [11, 36, 37, 49, 80]. While large language models (LLMs) have increasingly matched or even surpassed human capabilities in linguistic fluency, humans still hold a clear advantage in spatial reasoning over both language and multimodal models. Human spatial reasoning encompasses the ability to mentally manipulate objects, navigate complex environments, and critically—visualize abstract relationships [50, 88]. These capabilities, particularly the capacity to use spatial metaphors for non-spatial concepts and optimize within multi-constraint spaces, remain challenging for current AI systems despite their linguistic sophistication [22, 23]. This asymmetry motivates the design of mixed-initiative systems that combine human spatial reasoning with the linguistic intelligence of LLMs [21] to support complex cognitive tasks that require both sophisticated language generation and multi-dimensional reasoning, such as scientific writing that balances accuracy with accessibility across diverse audiences.

One notable form of spatial reasoning is direct manipulation of LLM output [81]: continuous feedback, rapid, reversible adjustments make complex intents expressible beyond text prompts alone. Systems such as ForceSPIRE [28] and Drag-and-Track [68] harness spatial operations to steer semantic analysis and data processing, bridging tacit goals and algorithmic execution. In LLM contexts, node-link diagrams [22] support GenAI-assisted hypothesis exploration, while real-time pipeline steering systems such as WaitGPT [96] enable fine-grained control over LLM workflows through spatial interactions. While direct manipulation interfaces are effective for illustrating step-by-step LLM processes, such sequential layouts quickly become cluttered as task complexity increases, limiting their ability to capture higher-level rhetorical trade-offs. To address this, researchers have turned to graph and tree-based views—such as Sensecape [84], Luminate [83], and Graphologue [45], which reveal relationships among generative elements via node–edge structures and support hierarchical exploration with LLM text output. Yet these systems largely prioritize inspection over in-situ steering of trade-offs.

As generative systems have demonstrated stronger linguistic capabilities, researchers have begun developing mixed-initiative visualization systems that combine human spatial reasoning with the linguistic intelligence of LLMs for collaborative text creation. For instance, sketch-driven storytelling interfaces allow users to spatially outline narrative trajectories, which are then expanded by language models into full-fledged text, thereby translating between spatial reasoning and linguistic generation [13]. Likewise, PatchView's "dust-and-magnet" metaphor enables users to rapidly cluster and combine narrative fragments through spatial manipulation [14], while Toyteller [15] transforms story fragments into interactive "toys" that encourage expressive ideation.

While these spatial approaches to human-AI collaboration have shown promise in creative domains and general text manipulation, their application to the specialized demands of science communication writing remains largely unexplored. Unlike creative domains, where LLM outputs are not bound by strict requirements and primarily seek new insights, science communication writing imposes stricter constraints, such as maintaining a linguistic balance between scientific exposition and narrative engagement [7, 16, 29, 78]. This gap represents a significant opportunity, as science communication writing inherently lends itself to spatial reasoning—writers naturally conceptualize their work through spatial metaphors such as "moving toward" accessibility, finding the "sweet spot" between detail and clarity [17], or "navigating" competing audience needs [98]. This research gap motivates us to design a 2D visualization interface that combines human spatial reasoning capabilities with LLM linguistic intelligence to support the iterative revision process of balancing scientific exposition and narrative engagement.

3 System Design

Based on our literature review, narrative engagement and scientific exposition are two critical dimensions that require careful consideration and when creating science communication narratives [25, 38, 62]. Writers must navigate an iterative, non-linear revision process as they continuously shuttle between these competing demands, often finding that improvements in one dimension can inadvertently compromise the other [29, 33, 69]. This creates a persistent struggle where writers lack systematic guidance for simultaneously optimizing both dimensions during their multi-pass revision workflow, leading to inefficient trial-and-error approaches that may favor one dimension at the expense of scientific exposition or reader engagement [64]. To understand how these two aspects are considered and how a balance is achieved in authentic creative processes, we conducted further expert interviews (Section 3.1) and a literature review (Section 3.3) to establish a more instructive guideline.

3.1 Formative Study

To better understand the workflows, goals, and tool needs of science communicators, we conducted in-depth interviews with four professionals: a TikTok science animator (20K+ followers), a YouTuber (10K+ subscribers), a science columnist on a Q&A platform (200K+ followers), and an educational video producer. Each interview lasted approximately 90 minutes and focused on three areas: (1) their typical content creation workflow, (2) how they balance communicative goals, and (3) how they use LLM tools in practice. The qualitative findings are as follows:

(1) The Core Challenge: Balancing Scientific Exposition and Narrative Engagement. Participants described two common workflows in science communication. The knowledge-to-stories approach, favored by those creating platform-independent or long-form content, begins with scientific concepts and adds narrative elements (e.g., examples, metaphors, stories) to enhance engagement. In contrast, the news-to-theories workflow—more typical of real-time or event-driven content—starts with current events or relatable experiences and layers in relevant scientific explanations. Despite differing starting points, all participants emphasized the same challenge: sustaining both scientific rigor and audience interest. One author noted, "If it's too technical, people stop watching. If it's too entertaining, they call it shallow." Across formats, authors stressed the need to balance clarity, credibility, and emotional connection.

(2) Narrative Strategies Are Essential but Lack Structured Support. To make their writing more engaging, participants reported deliberately applying narrative strategies, such as metaphors, real-world analogies, quotations, and personal anecdotes, to enhance the appeal of their content. One author revised content by adding narrative "hooks" after drafting the science explanation; another explicitly mapped theories to familiar experiences. The science columnist also said she relied on LLMs to quickly associate trending news with relevant theories. However, these four experts also noted that these decisions were largely intuitive due to their extensive editing and revision of texts and lacked structured support. They mentioned that it would be better to have a holistic narrative framework to guide the revision process. Additionally, they expressed a desire for clearer feedback on how well their narrative choices aligned with real audience feedback.

(3) LLMs Enable Exploration but Require Human Filtering for Precision. All four participants had experimented with LLMs to support writing, primarily for idea generation, tone adjustment, and connecting scientific ideas to familiar concepts. For example, the educator used LLMs to make explanations "more relaxed and child-friendly," while the columnist relied on them to quickly associate trending news with relevant theories. The YouTuber, who typically starts with expository theories, used LLMs to generate more examples and metaphors and edit based on the output to aid audience understanding. All four of them mentioned that co-creating with LLMs enabled them to revise content more quickly. They also noted that LLMs provided more examples and diverse perspectives to enhance the content's engagement and understanding, or to strengthen its scientific rigor and support. For example, the science columnist noted that she typically asks the LLM to surface a wide range of relevant theories, then filters through these options herself, and once one is selected, she carries out more fine-grained refinements. This illustrates how LLMs contribute linguistic intelligence by supporting both flexible exploration by surfacing diverse theoretical possibilities and fine-grained modification of specific content once a direction is chosen.

(4) Iterative Revision Relies on Intuition Due to Lack of Timely Feedback. Participants consistently emphasized that science communication writing is a highly iterative and non-linear process. They often went through multiple rounds of revision: starting with a draft focused on scientific explanation, then adding narrative elements, and finally refining language and visual expression. Each round could strengthen one dimension while weakening another. For instance, a YouTuber noted that after polishing the scientific argument, the storytelling often felt less engaging, requiring the addition of analogies or examples; yet when more narrative elements were included, there was concern that the content might lose academic rigor. These revisions were guided largely by intuition rather than systematic criteria. Audience feedback (e.g., views, likes, comments) was delayed, indirect, and rarely pinpointed which changes improved clarity or engagement. As the TikTok science animator noted, "You only know if it worked after publishing—and by then, it's too late." This lack of timely, fine-grained feedback left creators relying on trial-and-error, making it difficult to efficiently balance narrative appeal with scientific rigor.

In sum, science communicators need assitant to help them balance rigor and engagement, apply narrative strategies systematically, harness LLMs for exploration and refinement, and receive timely feedback. Addressing these needs would enable more efficient, intentional revision processes.

3.2 Design Goals

Drawing from the findings of the existing literature on science communication, as well as pilot testing on initial prototype and expert interview, we have established the following design goals:

Design Goal 1: Use Spatial Balancing to Visualize Trade-offs between Exposition and Narrative Engagement in Science Communication Writing. Prior work highlights the need to balance accurate exposition with engaging storytelling in science communication [25, 38, 62], and our formative interviews (Section 3.1) confirm that authors struggle to manage this tension. Writers often face implicit trade-offs—risking drafts that lean too heavily toward exposition or narrative—yet these shifts are difficult to track at the local level. To address this, the system should make both dimensions visible, helping authors evaluate relative levels of exposition and engagement without cognitive overload.

Design Goal 2: Guide Revisions with Strategy Scaffolds. Prior literature documents many techniques to address distinct communication objectives (see Section 3.3). Yet, LLM usage often requires authors to manually break down tasks and design prompts, which can be demanding [82].The system should therefore scaffold trategies—offering prompts, labels, etc. that help authors systematically select and apply approaches best suited to their communication goals. This reduces the burden of recalling strategies and allows for more deliberate, goal-oriented writing process.

Design Goal 3: Enable Flexible Exploration and Granular Controls Through Multi-Version Revision. Prior work [3, 65, 95, 98, 99] and our formative interviews show that in iterative revision, science communicator often relies on LLMs to explore multiple possibilities in pursuit of a specific goal, and then to perform fine-grained modifications within the selected direction. Yet effective writing frequently arises from exploring multiple possibilities through iterative drafting and deep refinement of specifci version [26]. Thus, the system should therefore support multi-version revision with non-linear history tracking and granular editing controls, enabling authors to revisit, merge, or revert drafts flexibly.

Design Goal 4: Embed Reflection Within Iterations to Support Self-Monitoring. Effective science communication with LLMs requires not only generating content but also iteratively revising and evaluating drafts with feedback [49]. Our formative study further underscored that authors receive little timely, fine-grained feedback during revision, leaving them to rely largely on intuition. To address this gap, the system should integrate lightweight reflection cues (e.g., visual indicators or checkpoints) into the workflow, prompting authors to pause, assess, and recalibrate. These signals help writers stay aligned with their goals and maintain control of the revision process.

Table 1. Labels of Science Communication Writing Strategies.

Scientific Exposition			
Label 1 **Articulate Precisely** Communicates scientific concepts with exposition and clarity, using appropriate terminology and well-defined language to prevent ambiguity or misinterpretation [44, 48, 64].	**Label 2** **Elaborate Thoroughly** Provides sufficient detail or comprehensive theoretical discussion by unpacking underlying mechanisms, explaining implications, and citing evidence to elaborate on the knowledge point while avoiding bias [52, 69].	**Label 3** **Verify Knowledge** Supports claims with credible sources, data, or reasoning, allowing audiences to feel more trustworthy of the given information [52, 73].	**Label 4** **Maintain Logical Consistency** Ensures that arguments and explanations are coherent and internally consistent, following a clear logical structure [89].
Strategies: (4) Acknowledge Uncertainties, (5) Consistent Terminology, (18) Simplify and abstract language, (19) Clarify Key Terms, (21) Repeat key point(s) or question(s), (22) Emphasize with Numbers	**Strategies:** (3) Step-by-Step Explanation, (4) Acknowledge Uncertainties, (7) Everyday Events to Scientific Insights, (22) Emphasize with Numbers, (25) Tie Science to Current Events	**Strategies:** (2) Rigorous Source Verification, (6) Citations & Quotes, (7) Everyday Events to Scientific Insights, (22) Emphasize with Numbers, (7) Everyday Events to Scientific Insights Events	**Strategies:** (1) Layered Transitions, (3) Step-by-Step Explanation, (20) Key Point Recap, (23) Strengthen the Connections Between Content
Narrative Engagement			
Label 5 **Captivate & Immerse** Engages the audience's attention and draws them into the narrative or content flow by adding stories [38, 57] or using intriguing language [29, 64].	**Label 6** **Enhance Understanding** Help audiences to grasp complex scientific ideas using rational, structural content or vivid analogies, visualizations [29, 38, 43].	**Label 7** **Inspire Curiosity** Stimulates the audience's desire to learn more and have motivation to further explore by applying different forms of questions [53].	**Label 8** **Evoke Emotion** Creates an emotional response, positive or negative, and makes the audience feel connected to the content, even immerse themselves in the described scenario [38, 74].
Strategies: (8) Question-Answer Hook, (9) Reflection Question, (10) Suspense-Driven Reveal, (11) Use metaphors, (12) Inject humor, (13) Add real-world supporting examples, (14) Add stories, (15) Add an imagery description, (16) Create negative emphasis for focused attention, (17) Make positive emotion to expand action repertoire	**Strategies:** (11) Use metaphors, (13) Add real-world supporting examples, (14) Add stories, (15) Add an imagery description, (21) Repeat key point(s) or question(s), (23) Strengthen the Connections Between Content, (24) Present Balanced Views, (25) Tie Science to Current Events	**Strategies:** (8) Question-Answer Hook, (9) Reflection Question, (10) Suspense-Driven Reveal	**Strategies:** (9) Reflection Question, (12) Inject humor, (14) Add stories, (16) Create negative emphasis for focused attention, (17) Make positive emotion to expand action repertoire, (21) Repeat key point(s) or question(s)

Note. Specific information about each strategy (e.g., definitions, examples) is presented in Table 4.

3.3 Strategies for Science Communication Narrative Design

Based on the results from the pilot interviews, we conducted a literature review in related fields, specifically in communication studies, education, psychology, linguistics and writing, and HCI, to identify writing strategies that can enhance narrative engagement and scientific exposition. We searched keywords "science communication" OR "scientific writing" OR "popular science" AND "strategy" OR "strategies" OR "method" in Google Scholar, the ACM Digital Library, and the IEEE Xplore Digital Library. After screening the abstract and full paper, we selected 47 papers, across Education (N=5), Psychology (N=7), Communication Studies (N=27), Linguistics and Writing (N=4), and HCI (N=6). We identified a total of 25 strategies from these selected papers. By using open coding [41] and design space analysis [10] methods, two authors developed and organized a design space (Table 4).

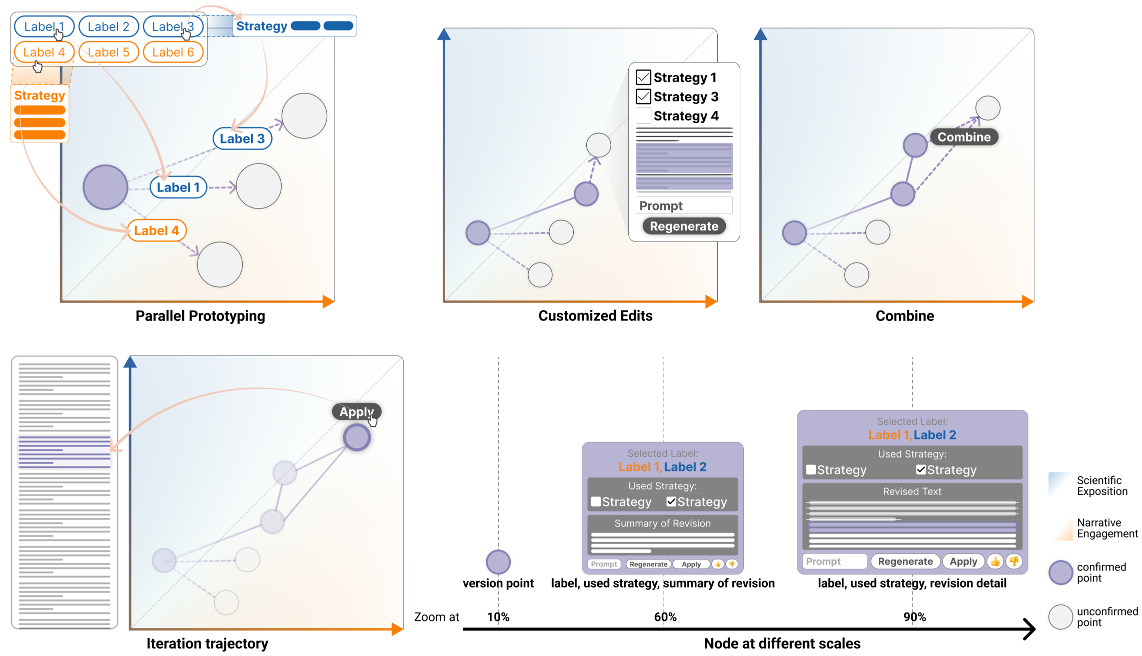

Fig. 2. (1) SpatialBalancing support parallel prototyping with diverse directions of LLM output; Authors can use customized edits like change specifc strategy and combine different LLM output to generate new nodes. The 2D coordinate space also allow author to see their iteration trajectory. (2) SpatialBalancing canvas supports three zoom levels: dots for version overview (0–30%), change summaries with labels and strategies (40–70%), and full content with highlights of edits (80–100%).

In this design space, we categorized the 25 identified strategies into three groups: those that enhance narrative engagement (N=10), those that enhance scientific exposition (N=7), and those that enhance both (N=8). Then, we conducted a Focus Group Discussion (FGD) [71] with the four experts. Together, we refined our initial strategy design space by clarifying the definition and use of each strategy, and classified the communication strategies by their functions. This process yielded four labels each for scientific exposition and narrative engagement. Some strategies, due to their multifunctionality, were assigned to multiple labels, forming the final design space (Table 1).

The defined strategies and their usage will serve as a prompts library for LLMs to support strategy selection and modification, while the corresponding examples will be applied in few-shot learning (Section 3.5.1).

3.4 Interface Design

3.4.1 SpatialBalancing Overview. SpatialBalancing comprises a left-hand text editor and a right-hand exploratory canvas (Figure 3). Authors can send any span—sentence, paragraph, or full draft—to the canvas for iterative revision. Each version is plotted in a 2D space (x: Narrative Engagement; y: Scientific Exposition); gray points denote exploratory drafts and purple points mark confirmed selections, which can be further refined via labels or custom edits. This spatial view makes revision states and decision points explicit, helping authors balance exposition and engagement.

The canvas supports branch-based exploration with three zoom levels (Figure 2). Dropped text becomes a root node; applying labels or custom instructions spawns child nodes, forming a tree that traces exploration paths. At 0–30% zoom, points provide an overview; at 40–70%, summaries show per-version changes and chosen strategies; at 80–100%, full text with diffs against the original is displayed. This progressive disclosure enables rapid comparison and reflective choice among alternatives.

Real-time Two-Axis Feedback (DG1 & DG4). Based on insights from metacognitive research, authors benefit from explicit feedback that reduces the cognitive burden of juggling multiple objectives (DG1) and allows self-monitoring of revision progress and alignment with writing intention (DG4). In SpatialBalancing, each version of the text is plotted as a point in a two-dimensional space, with one axis representing *narrative engagement* and the other *scientific exposition*. A "Scorer Agent," trained on audience ratings, assigns scores whenever authors drag a new piece of text into the canvas to create a node or perform additional edits that generate additional nodes. These scores determine the position of each node on the coordinate axes. By projecting revisions into a two-dimensional semantic space, the system externalizes abstract trade-offs into spatial patterns, supporting human spatial reasoning to quickly perceive balance. Meanwhile, the LLM-based Scorer Agent provides the linguistic intelligence to interpret audience-rated dimensions (engagement, exposition) and translate them into scores.

3.4.2 *Strategy Recommendation via Eight Labels (DG1 & DG2).*

(Figure 4 (1)) To support DG1 (reducing cognitive load) and DG2 (scaffolding revision), SpatialBalancing provides an eight-label taxonomy that represents core revision goals (e.g., inspire curiosity, elaborate thoroughly). Derived from expert interviews and literature, four labels target scientific exposition and four enhance narrative engagement. Users can select labels aligned with their revision intentions, while

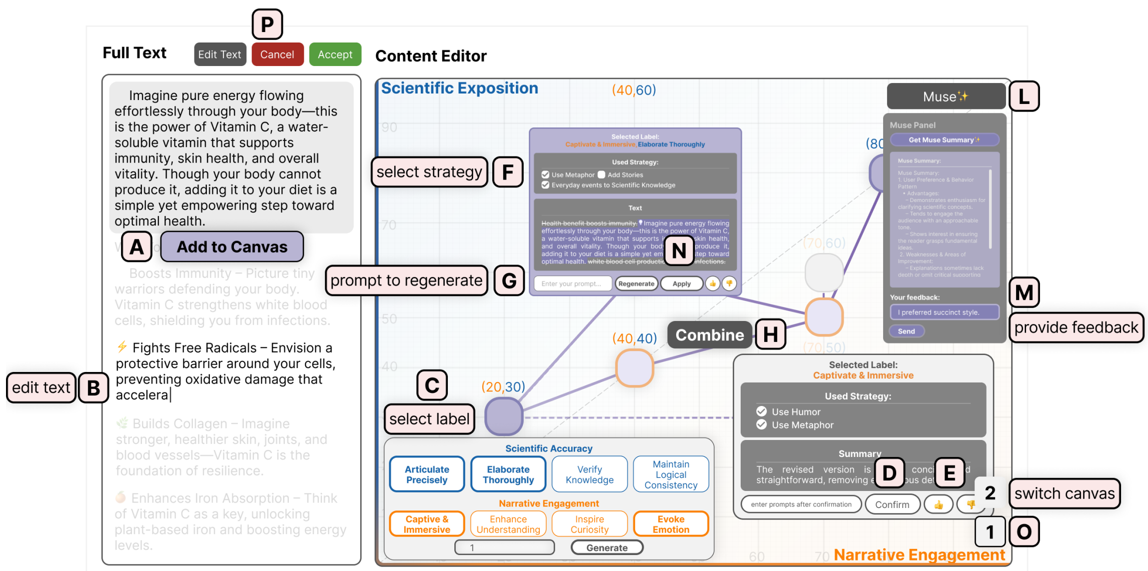

Fig. 3. The SpatialBalancing interface has two main sections: a text editor on the left for placing and directly editing source text (B), and a canvas on the right for revising selected segments (A). In the center, a visualization tracks iteration scores across narrative engagement and scientific exposition for multiple LLM-generated versions. Once a segment is confirmed for revision, authors assign labels (C) that guide editing directions and generate revision nodes. Within each node, content can be refined by entering custom prompts (G), switching strategies (F), or combining strategies from different nodes (H). Edits can be applied (N) to update the original text and view the full article. Muse (L), in the canvas's top-right corner, provides an overview of revision history and accepts author feedback (M), which informs future strategy recommendations. Editing other article sections opens a new canvas; authors can switch between revision records via the control in the bottom-right corner (O).

the LLM automatically draws on appropriate combinations of strategies to generate corresponding modifications based on the design space from the literature review as a prompt engineering library (Section 4). This design reduces the burden of recalling all possible options while guiding authors toward systematic, goal-directed revisions. The eight-label taxonomy further externalizes diffuse linguistic strategies into discrete, spatially mappable choices: authors use spatial reasoning to navigate directions, while the LLM Recommender Agent leverages linguistic intelligence to transform abstract strategies into concrete textual variants.

3.4.3 Fine-Grained Control for Specific Versions (DG3). (Figure 4(2)) To support DG3, authors can refine individual nodes after exploring different branches. Once a node is confirmed, it turns purple while unconfirmed nodes remain gray, visually distinguishing revision states. Three fine-tuning operations are available: toggling previously applied strategies, providing customized prompts (e.g., "try a different metaphor" or "make this more concise"), and merging

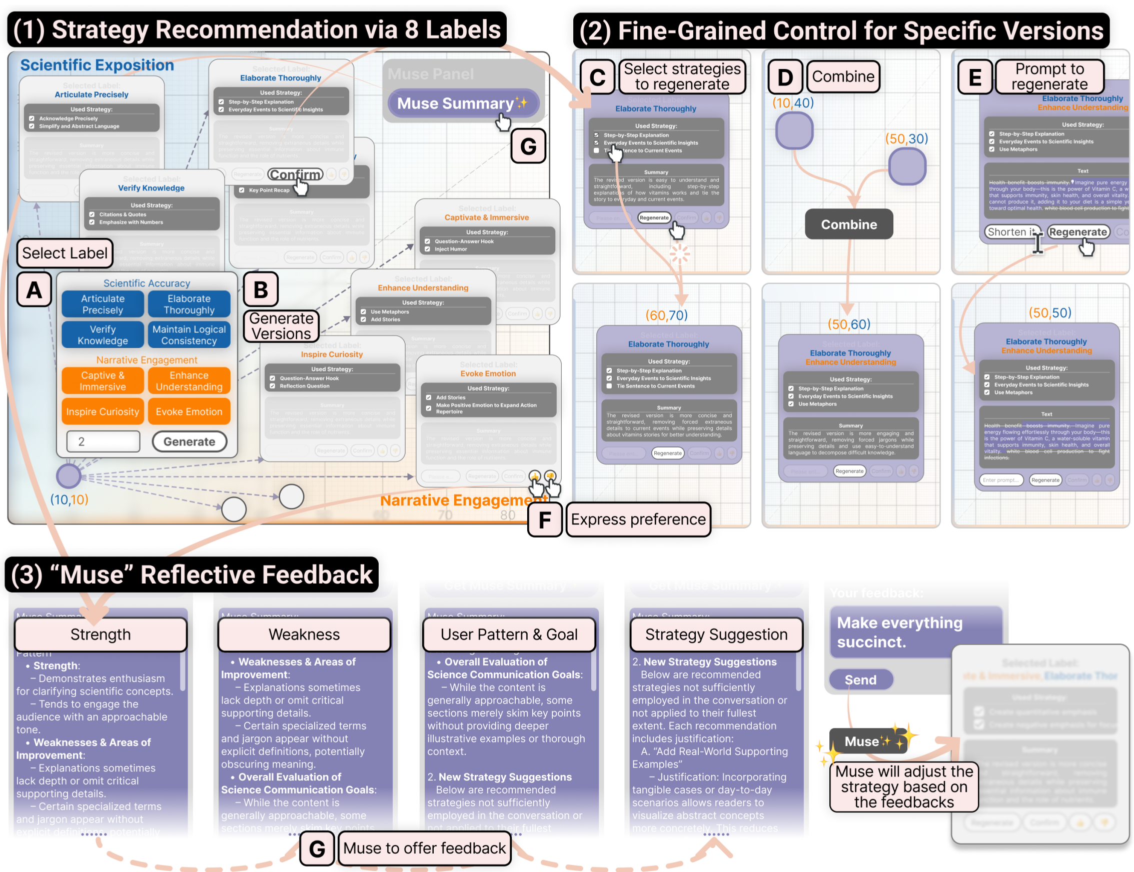

Fig. 4. (1) Strategy Recommendation via Eight Labels: SpatialBalancing offers eight revision labels—four enhancing narrative engagement and four strengthening scientific exposition. Authors can select one or more labels and specify the number of versions to generate under each; (2) Fine-Grained Control: Generated nodes can be refined by adjusting the applied strategies, merging nodes to combine labels, or entering custom prompts for tailored edits; (3) "Muse" Reflective Feedback: Muse provides iterative feedback on strengths, weaknesses, author patterns and goals, and strategy suggestions. Authors can endorse or reject this feedback, enabling the system to adapt future recommendations to their preferences.

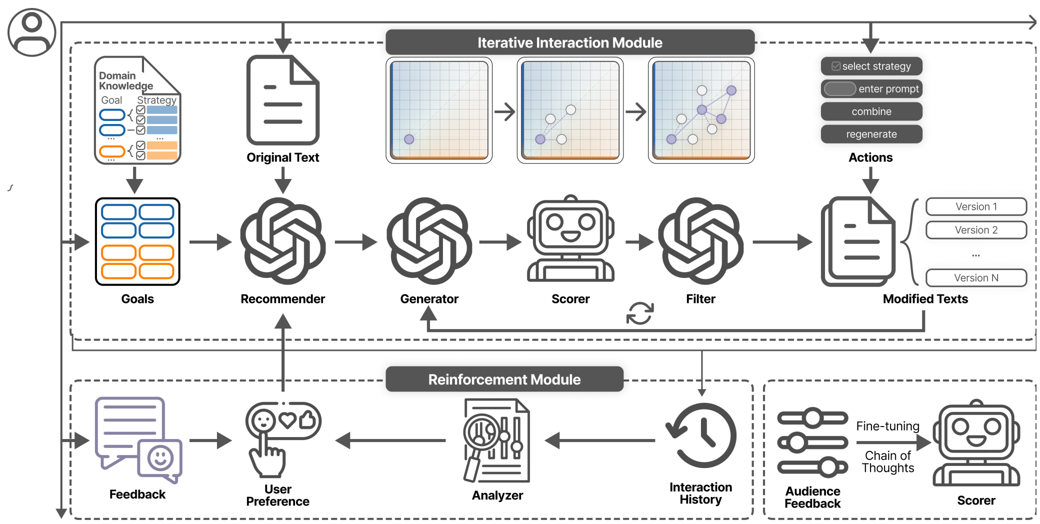

Fig. 5. SpatialBalancing backend overview. SpatialBalancing consists of two core modules: (1) The Iterative Interaction Module, where LLM-based agents—Recommender, Generator, Scorer, and Filter—collaboratively produce and evaluate multiple content versions based on narrative engagement and scientific exposition; and (2) the Reinforcement Module, which captures author feedback and inference based on interaction history of author behaviors to refine strategy recommendations through the Analyzer agent. This architecture supports adaptive text revision.

two versions to preserve strong elements from each. Visual branching and color cues engage human spatial reasoning to organize and differentiate versions, while LLM linguistic intelligence enables precise micro-level adaptations, grounding spatial manipulations in targeted linguistic outputs.

3.4.4 "Muse" Reflective Feedback (DG3 & DG4). (Figure 4(3)) To support DG3 and DG4, the Muse agent monitors author behaviors—such as node confirmations, strategy selections, and engagement–exposition choices—and synthesizes them into structured feedback. This feedback highlights strengths, weaknesses, editing patterns, and strategy suggestions, offering a clear channel for reflection. Authors can accept or reject suggestions, and their responses are fed back to the Recommender Agent to refine future recommendations. By integrating spatial activity traces with LLM-based linguistic analysis, Muse links behavioral patterns to tailored narrative and exposition strategies, enhancing self-awareness and promoting iterative refinement.

3.5 Backend and Implementation

The backend of SpatialBalancing comprises several LLM-based agents organized into two main modules: a generation module and a reinforcement module. The overall pipeline is in Figure 5.

3.5.1 Generation Module. This module begins by capturing the author's context and their selected modification labels. The system then proceeds into iterative processing handled by the following agents:

Recommender Agent: The recommender agent's core function is to generate multiple strategy combinations based on a author-selected label. When a author chooses a label, the agent analyzes the current textual features to identify the best combination from its associated strategy set (Section 3.3). Prompts are constructed using in-context learning and chain-of-thought principles based on the strategy design space (Table 4). The agent considers several factors when recommending strategies for each label, including strategy definitions, usage guides, examples, and the original text's role within the broader context of the entire text to recommend the most suitable strategies. The final output consists of multiple strategy combinations, which are then passed to the scorer to filter and select the top-scoring versions that has higher scientific exposition or narrative engagement score.

Generator Agent: The generator agent create child nodes based on author input instructions. When generating new content, the generator receives two types of input to form a new node: (1) strategy recommendations from the Recommender Agent, which are used to guide the generation of revised text that aligns with the author's chosen direction (Labels). The generator adopts in-context learning, referencing the recommended strategies' definitions, usage guidelines, and examples to perform content modifications based on the previous node (adopted from Section 3.3); and (2) author-specific refinements passed from the front end during regeneration. These refinements may include prompt adjustments, combining nodes, or deactivating particular strategies.

Scorer Agent: The scorer simulates real-time audience feedback by evaluating each generated version along two axes: Narrative Engagement (X) and Scientific Exposition (Y).

To support this, we curated a high-quality dataset of 45 science texts from five common science communication domain, varying in length and narrative style. Each text was revised by a science communication expert and annotated by 27 non-experts using a rubric developed by three domain experts. The rubric incorporated sub-dimensions of narrative engagement and scientific exposition (perceived credibility over strict factual correctness of the narrative). Scores were normalized to a 0–100 scale and used to fine-tune a GPT-4o model via a small-sample learning strategy[1]. This enables the scorer agent to give score to resemble human audience across both scientific exposition and narrative engagement. The scorer agent is powered by this fine-tuned GPT-4o model. Details on dataset construction and model training are provided in Appendix A.2.

To validate the reliability of the scoring mechanism, we conducted a technical evaluation comparing the accuracy of fine-tuned and non-fine-tuned scorers in simulating audience ratings. As shown in Table 2, the fine-tuned scorer exhibited much higher agreement with human ratings (r=0.90/0.91, RMSE≈6–7) than the non-fine-tuned model (r=0.84/0.57, RMSE=22–31). Detailed evaluation detail is provided in Appendix A.2.

Table 2. Evaluation of the similarity between fine-tuned and original GPT-4o models' scores and human scores.

Model	**Pearson Correlation**		**RMSE**	
	Engagement	Exposition	Engagement	Exposition
w/ FT	**0.90**	**0.91**	**6.48**	**7.02**
w/o FT	0.84	0.57	22.48	30.90

Filter Agent: This agent uses the scorer's outputs to select the top-k versions that best meet the author's expectations. Filter Agent ensures that the selected outputs not only fulfill the intended modification chosen direction (Labels) and achieve high scores but also filter out generated failures and low-quality content. This prevents content redundancy and enhances overall generation quality.

3.5.2 Reinforcement Module. Since author iterations form a tree of nodes enriched with valuable data (selected labels, prompts, likes /dislikes, and feedback), we developed an analyzer agent to harness both the explicit and implicit

[1]https://platform.openai.com/docs/guides/fine-tuning?utm_source=chatgpt.com

signals from these interactions. The analyzer agent captures behavioral data during the iterative process and uses chain-of-thought prompts to interpret author revision behavior.

Analyzer Agent: The analysis pursues two main goals: (1) identifying common editing patterns, including stylistic preferences, trade-offs between scientific exposition and narrative engagement, and individual author strengths or weaknesses; and (2) uncovering alternative or underused strategy directions. These insights are passed to the Muse component (Section 3.4.4). After the author provides feedback on the LLM's suggestions through Muse, the Analyzer Agent incorporates this real-time feedback (e.g., approvals or further edits) and updates the Recommender Agent accordingly. This process refines subsequent strategy recommendations, ensuring that each iteration aligns more closely with the author's preferences and habits. The feedback loop enables the system to adapt continuously to personal writing habits while balancing narrative engagement and scientific exposition throughout the revision process.

3.5.3 Implementation. SpatialBalancing is implemented as a web application, with a Python-based backend developed using Flask[2] framework and a frontend built using ReactFlow[3].

For the AI agents, we employ different LLMs tailored to their functional roles. The recommender, generator, and filter agents are powered by the GPT-4o-mini model, optimized for fast, high-quality content generation. The analyzer agent, which requires deeper reasoning to interpret author behavior and editing patterns, is supported by the GPT-o1 model—a reasoning-oriented LLM. For the scorer agent, it is powered by a fine-tuned GPT-4o model using a small-sample learning strategy[4]. The frontend into predefined prompt templates and communicates with the remote LLMs to obtain results. This modular design allows us to tailor agent behavior based on context while maintaining flexibility in prompt construction and LLM selection. The detailed use of prompts in the backend can be found in the Appendix A.7.

4 User Study

To further understand the effect of the SpatialBalancing system on users' experience during the science communication narrative writing process—particularly its impact on users' cognition and human-AI collaboration behavior patterns, we conducted a within-subjects user study involving 16 participants with prior experience in science communication. All participants were recruited from a local university. Each participant completed four text editing tasks: two using the SpatialBalancing system and two using a baseline system.

The baseline system used in this study was an interface consisting of a text editor and a conversational agent (powered by GPT-4o) that supported inline editing and suggestions from LLM. In both conditions, participants were provided with an Excel file containing a comprehensive strategy table. This table included the strategy name, definition, usage instructions, examples, and corresponding labels. Participants were encouraged to use this table as a reference and to copy-paste content into the prompt area as needed during the tasks.

4.1 Participants

We recruited 16 participants (9 male, 7 female; aged: 24-31 (M = 26.9, SD = 2.0)), all of whom held postgraduate degrees or higher. Most were PhD students, postdoctoral researchers, or university faculty members affiliated with a local university, possessing substantial experience in academic work, teaching, or public science communication.

Our system was not designed solely for expert science communicators but for a broad range of users with science communication needs, reflecting the growing diversity of science communication content creators in online platform [95,

[2]https://flask.palletsprojects.com/en/stable/
[3]https://github.com/wbkd/react-flow/
[4]https://platform.openai.com/docs/guides/fine-tuning?utm_source=chatgpt.com

98, 99]. Thus, participants varied in experience: 13 had hands-on practice in science communication (e.g., teaching undergraduates, producing explanatory media, or translating complex ideas), with six holding hybrid professional roles as creators, producers, journalists, or educators, while three primarily identified as consumers of science communication. LLM writing tool use also varied, with six using them daily, six weekly, and four occasionally. In terms of confidence, eight considered themselves strong science writers, while the other eight reported a more neutral stance, suggesting openness to support in expressing complex concepts for diverse audiences. The demographic information of these participants are in Appendix A.5.

4.2 Procedure

Each study session began with a live demonstration of the system. Participants were encouraged to explore the interface, try out features, and ask questions. During this walkthrough, the task objectives were also explained.

Each participant completed four text editing tasks: two using the SpatialBalancing system and two with the baseline. The texts were selected to represent two common styles of science communication: expository (e.g., "How mRNA Vaccines Work," "Criteria for Animal Domestication") and narrative storytelling (e.g., "Discovery of Archimedes' Principle," "Living and Thriving with ADHD"). Participants were asked to imagine two specific scenarios: (1) for the expository text: "I have a scientific narratives. How can I make it more engaging and interesting for an online science video?"; (2) for the narrative storytelling text: "I have a story as online science video narratives. How can I link it with more scientific concepts and add scientific credibility?" The length of each text averaged 297.75 words (SD = 19.64). The complete versions of the source texts used for the editing tasks are provided in Appendix A.3. To ensure balanced exposure and mitigate order effects or personal topic preferences, we counterbalanced both the system order (SpatialBalancing vs. baseline) and the text type assigned to each system. Thus, each participant edited one expository and one narrative text under each system condition.

Throughout the tasks, participants were encouraged to think aloud, verbalizing their thoughts, reasoning, and feelings as they interacted with the systems. All sessions were screen-recorded, and system interaction logs—such as button clicks (e.g., label selections, generate, regenerate, prompt input, combine)—were automatically captured for the SpatialBalancing condition.

4.3 Post-Task Survey and Instruments

After completing both conditions, participants completed a post-task survey with standardized instruments: the System Usability Scale (SUS) [6], NASA-TLX for workload [40], and the Creative Self-Efficacy Index (CSI) [12], with one item adapted to: "I think this system supported me in developing ideas or text collaboratively."

We also developed a concise co-creation survey targeting two metacognitive constructs from cognitive psychology [30, 79]. Metacognitive knowledge assessed awareness of cognitive goals (e.g., "I am aware of my writing goals during the editing process"). Metacognitive regulation captured planning, monitoring, and evaluation [70] (e.g., "I set specific goals for the narrative," "I reflect on editing strategies while using the AI tool," and "I reviewed the narrative to assess how well it communicated scientific content"). These items were adapted from the Metacognitive Awareness Inventory [79] and aligned with recent insights into AI-induced metacognitive demands. To measure perceived control during co-creation, we included items inspired by Human-AI interaction principles [90], focusing on participants' influence over outputs and narrative direction. Perceived autonomy was assessed according to Self-Determination Theory [20], addressing decision-making freedom, expressive latitude, and resistance to system pressure. The full list of items on metacognition, control, and autonomy is provided in Appendix A.4.

All instruments (NASA-TLX, SUS, CSI, and co-creation survey) employed a 7-point Likert scale. After task completion, each participant joined a 15-minute semi-structured interview designed to capture deeper insights into cognitive processes, feature usage, perceived system value, and moments of difficulty or breakthrough. These interviews complemented survey responses and enriched our understanding of user experience across both conditions.

5 Results

Our evaluation demonstrates that spatial reasoning serves as a powerful cognitive framework for managing the inherent tensions in science communication writing, transforming abstract balancing acts into concrete spatial navigation tasks. By externalizing the two-dimensional tradeoff between scientific exposition and narrative engagement through coordinate visualization, users developed enhanced spatial awareness of their revision choices, enabling them to treat writing quality not as a singular metric but as a navigable landscape with distinct directional goals. This spatial approach fundamentally shifted users' metacognitive processes, with participants showing significantly improved reflection on writing strategies ($M = 5.50$ vs. 4.63, $p = .013$) and strategic flexibility in adjusting approaches during editing ($M = 5.69$ vs. 4.56, $p = .016$) as they learned to "read" their position within the exposition-engagement space.

The spatial representation encouraged iterative exploration and balance-seeking behaviors, with users demonstrating significantly enhanced creative exploration ($M = 5.13$ vs. 3.69, $p = .004$) and increased enjoyment of the writing process ($M = 5.19$ vs. 4.13, $p = .039$) compared to traditional linear revision approaches. Remarkably, these cognitive and creative gains were achieved without imposing additional mental workload, as NASA-TLX results showed no significant differences across all six dimensions despite the system's expanded spatial reasoning capabilities. These findings reveal how spatial reasoning principles can be leveraged to scaffold complex writing decisions, enabling writers to develop more sophisticated mental models of quality that support both immediate revision choices and long-term strategic development.

5.1 RQ1: Spatial Reasoning in Science Communication Writing

5.1.1 2D Coordinate Visualization Facilitates Spatial Balancing for Informed Revision Decisions. The coordinate graph provides a persistent, actionable reference that maps abstract writing tradeoffs into tangible representation. Each node represents a version evaluated on two key dimensions: scientific exposition and narrative engagement. Most participants found the visualization facilitated revision prioritization. As P3 noted, "The coordinate graph is a feature that typical AI tools lack. It keeps me from getting lost balancing the two dimensions during revisions." Participants used scores to guide focus: P12 said, "I refer to the scores to decide which dimension I need to improve," while P6 observed, "If the two dimensions differ too much, it reminds me to pay attention to the other." By externalizing internal writing tradeoffs, the system facilitated metacognitive regulation through visualization of revision alignment and iteration comparison.

Besides, participants also used the graph to make informed revision decisions. P8 shared, "I can see strengths and weaknesses by comparing nodes; if scientific exposition drops, I adjust accordingly in the next generation." P10 added, "With the baseline, I had to judge on my own with no version comparison. Now I check if the engagement score is higher before reading carefully." The 2D coordinate space not only helps authors anticipate the direction of subsequent revisions but also enables them to compare and select among multiple versions based on their positions within the space. As P16 noted, "With multiple nodes, I can intuitively compare positions across dimensions, making differences clear and direct." Visual comparisons reinforced editorial confidence. As P3 explained, "Coordinate scores help me align edits with my standards and visually track progress; seeing engagement scores rise reinforces my decisions. It makes me feel that I am heading in the right direction."

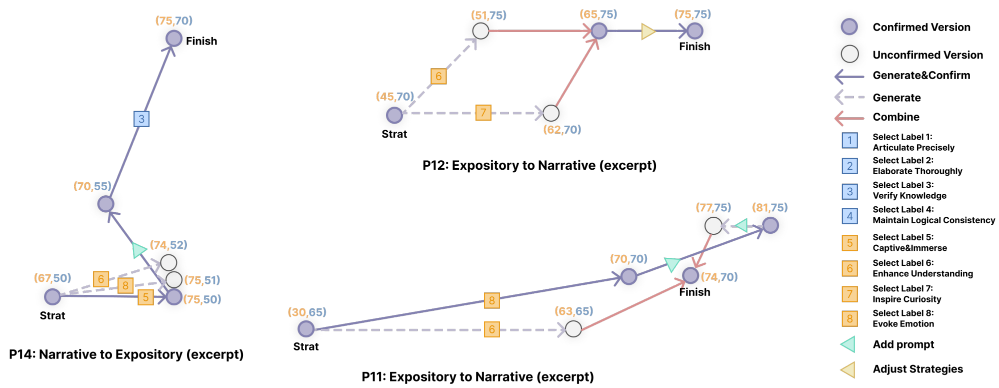

Fig. 6. Visualization examples of segment revisions from P11, P12, and P14.

In sum, the coordinate graph mapped scientific exposition and narrative engagement into a 2D space, helping authors compare versions and prioritize revisions. Participants used scores and positions to externalize tradeoffs, improving focus and efficiency. Visualization reinforced their sense of progress and boosted confidence in revision decisions.

5.1.2 Spatial Visualization Drives Iterative Balance-Seeking. The process of using the coordinate axes to assess current versions along the two dimensions constructively drove further iterations. As illustrated in Appendix A.6 (Figure 10), when attempting to add storytelling and narrative elements to expository content, participants initially selected labels associated with narrative engagement. However, during later iterations, they often returned to labels targeting scientific exposition in order to restore balance.

This kind of iteration can also be observed in Figure 6. For example, in the case of P14, when she attempted to revise a text from a narrative storytelling version to one with more scientific expression and explanatory content, she initially selected the label Captivate & Immerse, along with other engagement-enhancing labels. After fine-tuning the text at that stage using prompts, she realized the need to further improve scientific exposition. As a result, she selected the Verify Knowledge label and eventually accepted the final version.

This iterative back-and-forth highlights how spatial balancing supports users in dynamically regulating tradeoffs, ensuring their revisions move toward a more deliberate and well-aligned balance between exposition and engagement.

5.2 RQ2: Impact on Metacognitive Regulation and Creative Exploration

5.2.1 Enhanced Metacognitive Regulation and User Agency. To evaluate the system's impact on users' ability to reason about and adjust their writing strategies, we measured participants' reflection and adaptation while using SpatialBalancing to revise two articles from two directions. The results of metacognition, control, and autonomy are shown in Figure 7. SpatialBalancing received significantly higher ratings than the baseline on two dimensions: Q3- reflecting on one's own strategies ($M = 5.50$ vs. 4.63, $p = .013$) and Q4- adjusting strategies during the editing process ($M = 5.69$ vs. 4.56, $p = .016$). These results suggest that SpatialBalancing supports users in dynamically managing their writing strategies. For other dimensions, such as identifying areas for improvement, goal setting, and progress monitoring, SpatialBalancing also showed higher means.

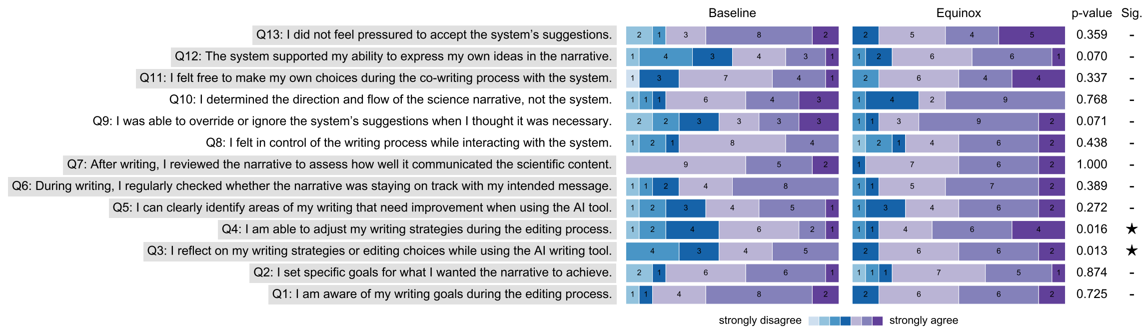

Fig. 7. Results of the Metacognition (Q1–Q7), Control (Q8–Q10), and Autonomy (Q11–Q13) questionnaires (p < .05 marked with *; p < .01 with **). Significant differences were observed in Metacognition: Q3 (M = 5.50 (SpatialBalancing) vs. 4.63 (Baseline), p = .013) and Q4 (M = 5.69 vs. 4.56, p = .016); marginal differences in Control: Q9 (M = 5.63 vs. 4.75, p = .071) and Autonomy: Q12 (M = 5.25 vs. 4.44, p = .070).

In terms of perceived control and autonomy, participants rated SpatialBalancing slightly higher across all items, especially in their ability to Q9- override system suggestions (M = 5.63 vs. 4.75, p = .071) and Q12- express their own ideas (M = 5.25 vs. 4.44, p = .070), although these did not reach significance. These trends indicate that SpatialBalancing fosters a stronger sense of authorship and agency in the LLM-supported writing process.

These findings indicate that SpatialBalancing enhances users' capacity for reflection and adaptation while reinforcing their role as active decision-makers. By supporting strategy calibration and the assertion of personal ideas, the system cultivates authorship and agency in the LLM-supported writing process.

5.2.2 Enabling Creativity Through Low-Cost, Flexible Exploration. The CSI questionnaire revealed that participants rated SpatialBalancing significantly higher in "Exploration" (M = 5.13 vs. 3.69 (Baseline), p = .004) and "Enjoyment" (M = 5.19 vs. 4.13, p = .039), indicating better support for exploring diverse narrative directions and enhanced writing experience. SpatialBalancing showed higher averages across all CSI items, demonstrating effective idea exploration without sacrificing usability (Figure 8).

Participants described interactions as playful and exploratory. P11 reflected, "I wanted to see how different strategies under the same label changed output, so I generated multiple versions. It gave me room to play and test." The system minimized cognitive overhead, enabling low-stakes, high-feedback interaction that encouraged curiosity.

The system provides flexibility for exploring multiple balancing directions while supporting fine-tuned adjustments within chosen axes. Unlike the baseline's linear process, this canvas-based interface facilitates parallel comparison and ongoing exploration. P6 noted, "These labels give me several options with different focuses simultaneously. I can choose one version to develop further and still return to earlier iterations after generating new branches." This non-linear workflow enabled reflective comparison without premature commitment.

The system occasionally catalyzed unexpected creativity. P11 recalled selecting "enhance understanding," which automatically inserted a metaphor: "That metaphor was so on-point, I hadn't even thought about that kind of revision before." Such moments illustrate potential for conceptual innovation beyond users' initial expectations.

Quantitative findings support this: participants rated SpatialBalancing higher for flexibility to "adjust writing strategies during editing" (M = 5.69 vs. 4.56, p = .016) (Figure 7 Q4) and exploration support for "diverse ideas and outcomes" (M =

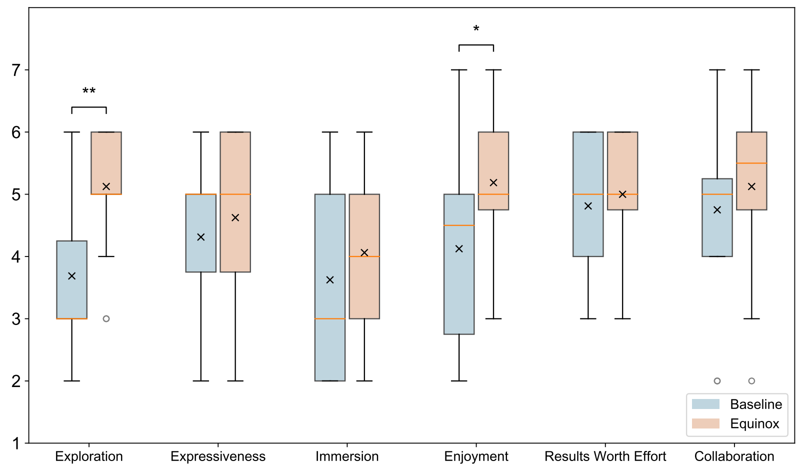

Fig. 8. The results of CSI questionnaire. ($*$: $p < 0.05$ and $**$: $p < 0.01$). Participants rated SpatialBalancing significantly higher in terms of "Exploration" (M = 5.13 (SpatialBalancing) vs. 3.69 (Baseline), p = .004) and "Enjoyment" (M = 5.19 vs. 4.13, p = .039)

5.13 vs. 3.69, p = .004) (Figure 8 Exploration). SpatialBalancing supports creativity by lowering experimentation costs, broadening revision possibilities, and enabling non-linear idea exploration.

Through playful interaction, flexible branching, and occasional novel rhetorical strategies, it encourages curiosity while maintaining user control, transforming revision from a constrained, linear task into an open-ended creative process.

5.2.3 Reflective Feedback through "Muse" Enhances Self-Awareness. Muse helps users recognize revision strengths and gaps through reflective feedback that mirrors their editorial process. P1 described a moment while revising Archimedes' principle: "A metaphor suggested by Muse struck me: buoyant force equals displaced water's weight, like balanced scale arms. This visual analogy illuminated the concept for me." Such feedback supports both evaluation and awareness of conceptual gaps.

The feedback prompted internalization of new strategies. P15 noted, "I started using strategies I hadn't tried before, and remembered to use them again." Several participants described how feedback reframed their broader writing approach. P6 said, "I started seeing where I tend to do well or poorly. Muse pointed out strengths I didn't even realize I had." P10 explained, "With more guidance during revision, I felt like I was internalizing a way of thinking. Even without the system, I'd know how to approach future writing."

This aligns with quantitative results showing SpatialBalancing better supports "reflecting on my writing strategies and choices" (M = 5.5 vs. 4.63 (Baseline), p = .013) (Figure 7 Q3). These results highlight that Muse's feedback fosters durable reflective habits that enhance self-awareness, strategic flexibility, and long-term writer development beyond immediate revisions.

5.3 RQ3: Interface Features' Contribution to Writing Quality and User Experience

5.3.1 Strategy Labels Enable Structured Exploration without Cognitive Overload. We evaluated cognitive workload and usability using NASA-TLX and SUS questionnaires (Table 3). NASA-TLX showed no significant differences between SpatialBalancing and baseline, indicating SpatialBalancing doesn't impose additional cognitive burden despite expanded

		SpatialBalancing		Baseline		Statistics	
		mean	std	mean	std	p-value	Sig.
NASA-TLX	Mental Demand	4.63	1.36	4.19	1.68	.404	—
	Physical Demand	3.19	1.60	2.63	0.96	.261	—
	Temporal Demand	2.63	1.36	3.19	1.38	.343	—
	Effort	3.94	1.39	4.44	1.79	.241	—
	Performance	5.13	0.89	4.88	0.96	.372	—
	Frustration	2.88	1.59	3.00	1.32	.724	—
SUS	Q1: use frequently	5.13	1.54	4.38	1.36	.155	—
	Q2: unnecessarily complex	3.00	1.41	2.94	0.85	.899	—
	Q3: easy to use	4.94	1.69	4.88	1.15	.964	—
	Q4: need support	3.94	1.91	2.81	1.87	.031	*
	Q5: function well integrated	5.13	1.26	3.44	1.36	.003	**
	Q6: inconsistency	3.06	1.39	3.25	1.53	.719	—
	Q7: learn to use quickly	4.88	1.59	5.06	1.44	.604	—
	Q8: awkward	2.44	1.26	2.50	1.37	.927	—
	Q9: confident	4.50	1.32	4.50	1.37	.812	—
	Q10: need learning	3.81	1.56	3.38	1.89	.397	—
	Overall Score	70.78	29.70	68.44	26.94	.729	—

Table 3. The statistical results of NASA-TLX and SUS questionnaires. (*: $p < 0.05$ and **: $p < 0.01$).

features. SUS revealed SpatialBalancing was more functionally integrated (Q5, p = .003) but required more user support (Q4, p = .031), suggesting richer capabilities with a learning curve. Overall usability scores were comparable: SpatialBalancing (M = 70.78) vs. baseline (M = 68.44).

According to participants(P1, P3, P6), this may be structured labels's ability to support strategy awareness and goal-oriented control by decomposing abstract objectives into manageable steps. Labels provide clear guidance and reduce the effort required for strategy knowledge retrieval, transforming ambiguous tasks into navigable concrete actions. User feedback revealed that labels not only improved execution efficiency ("strategies are packaged and I just click and go"(P7)) but also encouraged breaking habitual patterns and exploring new editorial approaches ("it gave me methods I hadn't considered"(P12)). Overall, the system maintains comparable usability while offering enhanced functionality through structured scaffolding, demonstrating that thoughtful interface design can expand users' capabilities without increasing cognitive load.

5.3.2 The Tension Between Guidance and User Judgment. Participants described how the system's visual and scoring feedback may influence their evaluation practices in subtle ways. While the coordinate axis enabled intuitive comparisons between revisions, some participants noted that the visibility and immediacy of scores could reduce their depth of textual engagement. As P4 reflected, "I outsourced a large part of the thinking process to the AI. It's faster and more efficient, but I also tend to think less carefully about the output as I trust the score results more than I did with the baseline."

Others expressed a degree of caution about over-relying on the scores. P16 noted that while the visual feedback was useful, "the scores are indicative rather than definitive. They sometimes do not reflect the actual quality of the generation and still require human judgment." Concerns about the interpretability of scoring were also raised. As P14 said, "Sometimes I don't know what an increase in score actually means. I can't tell whether each label contributes differently to the score or what specific content led to a higher score. I want to understand the logic behind the numbers."

These reflections suggest a potential tension: while the system offers accessible and actionable feedback, its effectiveness depends on users' ability to critically interpret the signals rather than accept them at face value. The interpretability of the scores also needs to be improved, as indicated by some participants.

5.3.3 Experienced Writers Seek More Flexible and Customizable Labels. While the fixed label set was seen as a helpful starting point, some experienced users felt it could be expanded to better support their advanced needs. P3, a seasoned science communicator, shared: "The eight labels are a solid foundation, but I would appreciate a broader set to support more diverse explorations." P1, P3, P2, and P14, all of whom are experienced science communicators or experienced writers, expressed interest in more customizable labels, such as they can combining or tailoring underlying strategies to form customized labels to align more closely with their specific goals. P14 also noted, "In addition to the current style-focused labels, it would be helpful to include others that target areas in writing revision like grammar or tone." This indicates a demand for labels that can be tailored to individual needs.

5.3.4 Muse as a Future Co-Editor. While participants appreciated what Muse could already do, many imagined what it might become. P2 wanted more real-time dialogue: "I wish it were more interactive—like chatting with someone who helps me reflect as I go." P14 hoped for more adaptability: "The more I use it, the more I want it to understand how I write and suggest things based on that." Others wished for more precision in the feedback. "Right now, Muse gives high-level suggestions," one participant said. "But it'd be more useful if it could point to which step or decision was strong or weak, and explain why." These comments suggest that participants saw Muse not just as a tool for generating or revising text, but as a partner that could grow with them—learning their writing style, giving relevant feedback, and helping them refine how they think through revisions.

6 Discussion

6.1 Designing Mixed-Initiative Human–AI Collaboration through Spatial Reasoning and LLM Linguistic Intelligence: Insights and Implications for Future Systems

LLMs have increasingly approached—and in some cases surpassed—human capabilities in generating fluent and diverse text. Yet despite these advances, LLMs remain limited in managing multi-objective trade-offs and comprehending abstract structures. Humans, in contrast, excel at spatial reasoning: visualizing abstract relationships, navigating multidimensional spaces, and balancing competing goals holistically. This asymmetry motivates combining complementary strengths so that humans remain active shapers of communicative outcomes rather than passive recipients of machine-generated text. Notably, just three years ago researchers were exploring transforming visual sketches into stories [13]; today, leveraging spatial reasoning to harness LLM linguistic intelligence has shifted from a desirable option to an essential capability.

Building on this motivation, our evaluation shows that integrating spatial reasoning with LLM linguistic capabilities turns complex writing decisions from abstract balancing acts into concrete navigational tasks. (1) By externalizing the two-dimensional trade-off between scientific exposition and narrative engagement via coordinate visualization, users employed spatial cognition to rapidly evaluate LLM outputs using positions as a reference beyond textual reading, increasing confidence in co-directing revision trajectories. (2) This spatial–linguistic integration yielded significant benefits: users leveraged iterative coordinate trajectories to better understand content development, exercised stronger control over LLM collaboration, and performed parallel comparison of multiple LLM outputs by spatial positioning to make better judgments—supporting richer exploration and greater enjoyment.

Following Tankelevitch et al.'s [86] dual-path framework, SpatialBalancing supports metacognition in two complementary ways. First, it *enhances* abilities by translating revision goals into spatial waypoints that scaffold planning (set targets), monitoring (track position/trajectory), and strategic control (retarget direction); strategy labels render intent explicit with predictable effects [94]. Second, it *reduces* metacognitive demand by shifting evaluative effort to ambient spatial cues: heuristic zones offer stopping rules for confidence calibration, and lattice/zoom views enable quick screening-to-detail transitions with lower working-memory load, yielding efficient, low-friction judgments [85].

Situating these contributions within related research, our spatial–linguistic approach complements *direct-manipulation and node-based systems* that emphasize stepwise control (e.g., ForceSPIRE [28]; Drag-and-Track [68]; WaitGPT [96]) as well as *graph- and tree-inspection approaches* (e.g., Sensecape [84]; Luminate [83]). It further extends *sketch- and fragment-driven spatial storytelling tools* (e.g., PatchView [14]; Toyteller [15]) by directly mapping rhetorical goals—exposition versus engagement—into a navigable space for in-situ steering of linguistic outputs across scales (from macro narrative to micro style).

Together, these results highlight fundamental **design principles** for future mixed-initiative systems that integrate human spatial cognition with LLM linguistic capabilities.

(1) Spatial-guided linguistic generation: users should dynamically define evaluation axes that direct LLM text production toward specific rhetorical goals, enabling spatial positioning to inform corresponding LLM linguistic adjustments from macro-narrative shifts to micro-stylistic changes.
(2) Direct spatial-linguistic manipulation: interfaces should allow users to drag coordinate nodes to generate linguistically-targeted revisions, where spatial movements trigger corresponding linguistic transformations that match desired positional targets.
(3) Collaborative spatial orchestration: systems should enable multiple contributors to spatially coordinate LLM linguistic outputs across different regions, positioning human spatial intelligence as the steering mechanism for orchestrating LLM generative capabilities in shared authoring contexts.
(4) Adaptive scaffolding: systems should dynamically adjust spatial guidance based on task complexity and user expertise, transitioning from dense spatial cues for novices to sparse, configurable environments for experts. This prevents over-dependence on LLMs while fostering collaborative metacognitive partnerships.
(5) Metacognitive transparency: systems should make LLM reasoning processes spatially visible, enabling users to understand why certain regions are highlighted. Such transparency supports appropriate trust calibration and maintains critical evaluation skills, ensuring that human spatial intelligence and LLM capabilities mutually enhance rather than replace each other in complex decision-making.

6.2 Limitation and Future Work

We describe several limitations in the study to define the scope of our findings clearly and motivate future work.

6.2.1 Lack of Evaluation on Text Quality and Communication Effectiveness. One limitation of the current study is the absence of a systematic evaluation of the generated texts. While the system produces revised versions of scientific narratives, we did not assess whether these revisions lead to improvements in quality for science communication purposes. Future studies could investigate whether the generated texts are more engaging, whether they enhance the perceived exposition of the information, or whether they facilitate better knowledge retention among audiences. Objective and subjective measures, such as engagement metrics, audience feedback, and comprehension tests, could be employed to evaluate the effectiveness of the texts in real-world science communication settings.

6.2.2 *Evaluation Dependency on Proxy Scores.*

Although SpatialBalancing provides real-time feedback on scientific exposition and narrative engagement, this feedback is generated by a model trained on proxy metrics (e.g., perceived credibility and engagement from non-experts). While useful, these proxies may not fully capture the nuance of effectiveness in real-world science communication. Actual audience reactions in diverse contexts (e.g., classroom learning vs. YouTube videos) may differ from model predictions. Therefore, the reliability and generalizability of the scoring system should be validated further.

6.2.3 *Methodological Limitations.*

This work has common methodological limitations including the short-term nature of system testing which may not reveal long-term adoption patterns, and the relatively homogeneous participant demographics that may not represent all potential user groups. Future work will aim to address the previously mentioned and these limitations through more comprehensive evaluations.

7 Conclusion

We presented SpatialBalancing, a writing interface that harnesses human spatial reasoning to navigate LLM-generated revision options in science communication. By visualizing the trade-off between scientific exposition and narrative engagement in a dual-axis space, the system enables users to iteratively balance competing communicative goals through spatial navigation. Our study shows this approach enhances metacognitive regulation and creative exploration, demonstrating how coupling human spatial cognition with AI linguistic capabilities supports deliberate revision toward balanced science communication.

References

[1] J Craig Andrews and Terence A Shimp. 2018. *Advertising, promotion, and other aspects of integrated marketing communications.* Cengage Learning.

[2] Isabelle Augenstein. 2021. Determining the credibility of science communication. *arXiv preprint arXiv:2105.14473* (2021).

[3] Tal August, Lauren Kim, Katharina Reinecke, and Noah A Smith. 2020. Writing strategies for science communication: Data and computational analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* 5327–5344.

[4] Besma Boubertakh. 2015. Towards Further Experimental Reproducibility: Making A Balance Between Conciseness, Precision and Comprehensiveness in Scientific Communication. *Journal of Neurology and Stroke* 3, 1 (Oct. 2015). https://doi.org/10.15406/JNSK.2015.03.00077

[5] Peter Broks. 2006. *Understanding popular science.* McGraw-Hill Education (UK).

[6] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[7] Terry W Burns, D John O'Connor, and Susan M Stocklmayer. 2003. Science communication: a contemporary definition. *Public understanding of science* 12, 2 (2003), 183–202.

[8] Rick Busselle and Helena Bilandzic. 2009. Measuring narrative engagement. *Media psychology* 12, 4 (2009), 321–347.

[9] Rocco Caferra, Giuseppe Di Liddo, Andrea Morone, and David Stadelmann. 2025. The media morphosis of science communication during crises. *Scientific Reports* 15, 1 (2025), 5506.

[10] Stuart K Card, Jock D Mackinlay, and George G Robertson. 1991. A morphological analysis of the design space of input devices. *ACM Transactions on Information Systems (TOIS)* 9, 2 (1991), 99–122.

[11] Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2024. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web* 27, 4 (2024), 42.

[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] 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

[14] 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.

[15] John Joon Young Chung, Melissa Roemmele, and Max Kreminski. 2025. Toyteller: AI-powered Visual Storytelling Through Toy-Playing with Character Symbols. https://doi.org/10.1145/3706598.3713435 arXiv:2501.13284 [cs].

[16] Michael F Dahlstrom. 2014. Using narratives and storytelling to communicate science with nonexpert audiences. *Proceedings of the National Academy of Sciences* 111, Supplement 4 (2014), 13614–13620.

[17] Michael F. Dahlstrom. 2014. Using narratives and storytelling to communicate science with nonexpert audiences. *Proceedings of the National Academy of Sciences* 111, supplement_4 (2014), 13614–13620. https://doi.org/10.1073/pnas.1320645111 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1320645111

[18] Michael F. Dahlstrom and Dietram A. Scheufele. 2018. (Escaping) the paradox of scientific storytelling. *PLOS Biology* 16, 10 (10 2018), 1–4. https://doi.org/10.1371/journal.pbio.2006720

[19] Andreas W Daum. 2009. Varieties of popular science and the transformations of public knowledge: some historical reflections. *Isis* 100, 2 (2009), 319–332.

[20] Edward L Deci and Richard M Ryan. 2012. Self-determination theory. *Handbook of theories of social psychology* 1, 20 (2012), 416–436.

[21] Zijian Ding. 2024. Towards Intent-based User Interfaces: Charting the Design Space of Intent-AI Interactions Across Task Types. *arXiv preprint arXiv:2404.18196* (2024).

[22] Zijian Ding, Michelle Brachman, Joel Chan, and Werner Geyer. 2025. "The Diagram is like Guardrails": Structuring GenAI-assisted Hypotheses Exploration with an Interactive Shared Representation. (2025).

[23] Zijian Ding, Fenghai Li, Haofei Yu, and Joel Chan. [n. d.]. Towards Direct Intent Manipulation: Drag-Based Research Ideation, Evaluation and Evolution. ([n. d.]).

[24] Anne DiPardo. 1990. Narrative knowers, expository knowledge: Discourse as a dialectic. *Written communication* 7, 1 (1990), 59–95.

[25] Julie S. Downs. 2014. Prescriptive scientific narratives for communicating usable science. *Proceedings of the National Academy of Sciences* 111, supplement_4 (Sept. 2014), 13627–13633. https://doi.org/10.1073/pnas.1317502111 Publisher: Proceedings of the National Academy of Sciences.

[26] Grant Eckstein, Jessica Chariton, and Robb Mark McCollum. 2011. Multi-draft composing: An iterative model for academic argument writing. *Journal of English for academic purposes* 10, 3 (2011), 162–172.

[27] Lee Ellis. 2022. Improving Scientific Communication by Altering Citation and Referencing Methods. *Journal of Social Science Studies* 9, 1 (2022), 1–1. https://doi.org/10.5296/jsss.v9i1.19548

[28] Alex Endert, Patrick Fiaux, and Chris North. 2012. Semantic Interaction for Visual Text Analytics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 473–482. https://doi.org/10.1145/2207676.2207741

[29] Wiebke Finkler and Bienvenido León-Anguiano. 2019. The power of storytelling and video: a visual rhetoric for science communication. (2019).

[30] John H Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist* 34, 10 (1979), 906.

[31] Laura Fogg-Rogers, Ann Grand, and Margarida Sardo. 2015. Beyond dissemination - Science communication as impact. 14, 3 (Sept. 2015). https://doi.org/10.22323/2.14030301

[32] Barbara L Fredrickson. 2004. The broaden–and–build theory of positive emotions. *Philosophical transactions of the royal society of London. Series B: Biological Sciences* 359, 1449 (2004), 1367–1377.

[33] Eric L Garland, Barbara Fredrickson, Ann M Kring, David P Johnson, Piper S Meyer, and David L Penn. 2010. Upward spirals of positive emotions counter downward spirals of negativity: Insights from the broaden-and-build theory and affective neuroscience on the treatment of emotion dysfunctions and deficits in psychopathology. *Clinical psychology review* 30, 7 (2010), 849–864.

[34] Katy Ilonka Gero and Lydia B. Chilton. 2019. Metaphoria: An Algorithmic Companion for Metaphor Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300526

[35] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for Science Writing using Language Models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1002–1019. https://doi.org/10.1145/3532106.3533533

[36] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for science writing using language models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*. 1002–1019.

[37] Katy Ilonka Gero, Vivian Liu, Sarah Huang, Jennifer Lee, and Lydia B Chilton. 2021. What makes tweetorials tick: How experts communicate complex topics on twitter. *Proceedings of the ACM on Human-computer Interaction* 5, CSCW2 (2021), 1–26.

[38] Manuela Glaser, Bärbel Garsoffky, and Stephan Schwan. 2009. Narrative-based learning: Possible benefits and problems. (2009).

[39] Jean Goodwin and Michael F. Dahlstrom. 2014. Communication strategies for earning trust in climate change debates. *WIREs Climate Change* 5, 1 (2014), 151–160. https://doi.org/10.1002/wcc.262 arXiv:https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcc.262

[40] Sandra G Hart. 1986. NASA task load index (TLX). (1986).

[41] Judith A Holton. 2007. The coding process and its challenges. *The Sage handbook of grounded theory* 3 (2007), 265–289.

[42] Tianle Huang and Will J. Grant. 2020. A Good Story Well Told: Storytelling Components That Impact Science Video Popularity on YouTube. *Frontiers in Communication* 5 (Oct. 2020). https://doi.org/10.3389/fcomm.2020.581349 Publisher: Frontiers.

[43] Tianle Huang and Will J Grant. 2020. A good story well told: Storytelling components that impact science video popularity on YouTube. *Frontiers in Communication* 5 (2020), 86.

[44] Oksana Ivchenko and Natalia Grabar. 2022. Impact of the Text Simplification on Understanding. In *Challenges of Trustable AI and Added-Value on Health*. IOS Press, 634–638. https://doi.org/10.3233/SHTI220546

[45] Peiling Jiang, Jude Rayan, Steven P. Dow, and Haijun Xia. 2023. Graphologue: Exploring Large Language Model Responses with Interactive Diagrams. https://doi.org/10.1145/3586183.3606737 arXiv:2305.11473 [cs].

[46] Klemens Kappel and Sebastian Jon Holmen. 2019. Why science communication, and does it work? A taxonomy of science communication aims and a survey of the empirical evidence. *Frontiers in communication* 4 (2019), 55.

[47] Jagdish Kaur. 2012. Saying it again: enhancing clarity in English as a lingua franca (ELF) talk through self-repetition. *Text & Talk* 32, 5 (Jan. 2012), 593–613. https://doi.org/10.1515/TEXT-2012-0028

[48] Martin Kerwer, Anita Chasiotis, Johannes Stricker, Armin Günther, and Tom Rosman. 2021. Straight From the Scientist's Mouth—Plain Language Summaries Promote Laypeople's Comprehension and Knowledge Acquisition When Reading About Individual Research Findings in Psychology. *Collabra: Psychology* 7, 1 (02 2021), 18898. https://doi.org/10.1525/collabra.18898 arXiv:https://online.ucpress.edu/collabra/article-pdf/7/1/18898/835600/collabra_2021_7_1_18898.pdf

[49] Jeongyeon Kim, Sangho Suh, Lydia B Chilton, and Haijun Xia. 2023. Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 115–135.

[50] Markus Knauff. 2013. *Space to reason: A spatial theory of human thought*. Mit Press.

[51] Laura M König, Marlene S Altenmüller, Julian Fick, Jan Crusius, Oliver Genschow, and Melanie Sauerland. 2024. How to communicate science to the public? Recommendations for effective written communication derived from a systematic review. *Zeitschrift für Psychologie* (2024). https://doi.org/10.1027/2151-2604/a000572

[52] Christoph Kueffer and Brendon M. H. Larson. 2014. Responsible Use of Language in Scientific Writing and Science Communication. *BioScience* 64, 8 (06 2014), 719–724. https://doi.org/10.1093/biosci/biu084 arXiv:https://academic.oup.com/bioscience/article-pdf/64/8/719/8719054/biu084.pdf

[53] Joe Lambert. 2013. *Digital storytelling: Capturing lives, creating community*. Routledge.

[54] Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A. Alghamdi, Tal August, Avinash Bhat, Madiha Zahrah Choksi, Senjuti Dutta, Jin L.C. Guo, Md Naimul Hoque, Yewon Kim, Simon Knight, Seyed Parsa Neshaei, Antonette Shibani, Disha Shrivastava, Lila Shroff, Agnia Sergeyuk, Jessi Stark, Sarah Sterman, Sitong Wang, Antoine Bosselut, Daniel Buschek, Joseph Chee Chang, Sherol Chen, Max Kreminski, Joonsuk Park, Roy Pea, Eugenia Ha Rim Rho, Zejiang Shen, and Pao Siangliulue. 2024. A Design Space for Intelligent and Interactive Writing Assistants. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–35. https://doi.org/10.1145/3613904.3642697 rate: 4.

[55] Robert A Lehrman and Eric Schnure. 2019. *The Political Speechwriter's Companion: A Guide for Writers and Speakers*. CQ Press.

[56] Marcia C Linn. 2025. Emergence and Characterization of Sex Differences in Spatial Ability: A Meta-Analysis. (2025).

[57] Norma J Livo and Sandra A Rietz. 1986. Storytelling: Process and practice. *(No Title)* (1986).

[58] Tao Long, Katy Ilonka Gero, and Lydia B Chilton. 2024. Not Just Novelty: A Longitudinal Study on Utility and Customization of an AI Workflow. In *Designing Interactive Systems Conference*. ACM, IT University of Copenhagen Denmark, 782–803. https://doi.org/10.1145/3643834.3661587

[59] Tao Long, Dorothy Zhang, Grace Li, Batool Taraif, Samia Menon, Kynnedy Simone Smith, Sitong Wang, Katy Ilonka Gero, and Lydia B Chilton. 2023. Tweetorial hooks: generative AI tools to motivate science on social media. *arXiv preprint arXiv:2305.12265* (2023).

[60] Roger Maskill. 1988. Logical Language, Natural Strategies and the Teaching of Science. *International Journal of Science Education* 10, 5 (1988), 485–495. https://doi.org/10.1080/0950069880100502

[61] Damien Masson, Zixin Zhao, and Fanny Chevalier. 2025. Visual Story-Writing: Writing by Manipulating Visual Representations of Stories. (2025).

[62] Daniel Gary McDonald Daniel Gary McDonald. 2014. Narrative research in communication: key principles and issues. *Review of Communication Research* 2 (2014), 115–132.

[63] Julia Metag, Florian Wintterlin, and Kira Klinger. 2023. science communication in the digital age—new actors, environments, and practices. *Media and Communication* 11, 1 (2023), 212–216.

[64] Jesús Muñoz Morcillo, Klemens Czurda, and Caroline Trotha. 2016. Typologies of the popular science web video. *Journal of Science Communication* 15 (May 2016), A02. https://doi.org/10.22323/2.15040202

[65] Nathan Ni. [n. d.]. Building a Scientific Narrative. https://www.the-scientist.com/building-a-scientific-narrative-71780.

[66] National Academies of Sciences, Medicine, Division of Behavioral, Social Sciences, Committee on the Science of Science Communication, and A Research Agenda. 2017. Communicating science effectively: A research agenda. (2017).

[67] Reham Omar, Ishika Dhall, Panos Kalnis, and Essam Mansour. 2023. A universal question-answering platform for knowledge graphs. *Proceedings of the ACM on Management of Data* 1, 1 (2023), 1–25.

[68] Daniel Orban, Daniel F. Keefe, Ayan Biswas, James Ahrens, and David Rogers. 2019. Drag and Track: A Direct Manipulation Interface for Contextualizing Data Instances within a Continuous Parameter Space. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 256–266. https://doi.org/10.1109/TVCG.2018.2865051

[69] Roger A Pielke Jr. 2007. *The honest broker: Making sense of science in policy and politics*. Cambridge University Press.

[70] Chenghai Qin, Ruru Zhang, and Yanling Xiao. 2022. A questionnaire-based validation of metacognitive strategies in writing and their predictive effects on the writing performance of English as foreign language student writers. *Frontiers in Psychology* 13 (2022), 1071907.

[71] Fatemeh Rabiee. 2004. Focus-group interview and data analysis. *Proceedings of the nutrition society* 63, 4 (2004), 655–660.

[72] Marissa Radensky, Daniel S Weld, Joseph Chee Chang, Pao Siangliulue, and Jonathan Bragg. 2024. Let's Get to the Point: LLM-Supported Planning, Drafting, and Revising of Research-Paper Blog Posts. *arXiv preprint arXiv:2406.10370* (2024).

[73] Marie-Claude Roland. 2009. Quality and integrity in scientific writing: prerequisites for quality in science communication. *Journal of Science Communication* 8, 2 (2009), A04. https://doi.org/10.22323/2.08020204

[74] Gillian Rowe, Jacob B Hirsh, and Adam K Anderson. 2007. Positive affect increases the breadth of attentional selection. *Proceedings of the National Academy of Sciences* 104, 1 (2007), 383–388.

[75] Maximilian Roßmann. 2025. Science Correction as a Communication Problem: Insights from Four Theoretical Lenses. *OSF Preprints* (3 February 2025). https://doi.org/10.31219/osf.io/82duj_v3

[76] Margaret A Rubega, Kevin R Burgio, A Andrew M MacDonald, Anne Oeldorf-Hirsch, Robert S Capers, and Robert Wyss. 2021. Assessment by audiences shows little effect of science communication training. *Science Communication* 43, 2 (2021), 139–169.

[77] Carol D. Saunders, Amara T. Brook, and Olin Eugene Myers. 2006. Using Psychology to Save Biodiversity and Human Well-Being. *Conservation Biology* 20, 3 (2006), 702–705. http://www.jstor.org/stable/3879236

[78] Keegan Sawyer and Brooke Smith. 2024. Communication and engagement for basic science: insights and practical considerations. *Journal of Science Communication* 23, 7 (2024), Y01.

[79] Gregory Schraw and Rayne Sperling Dennison. 1994. Assessing metacognitive awareness. *Contemporary educational psychology* 19, 4 (1994), 460–475.

[80] Yijia Shao, Yucheng Jiang, Theodore A Kanell, Peter Xu, Omar Khattab, and Monica S Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. *arXiv preprint arXiv:2402.14207* (2024).

[81] Ben Shneiderman. 1982. The future of interactive systems and the emergence of direct manipulation. *Behaviour & Information Technology* 1, 3 (1982), 237–256.

[82] Hari Subramonyam, Roy Pea, Christopher Pondoc, Maneesh Agrawala, and Colleen Seifert. 2024. Bridging the gulf of envisioning: Cognitive challenges in prompt based interactions with LLMs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.

[83] 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 CHI Conference on Human Factors in Computing Systems*. 1–26.

[84] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM, San Francisco CA USA, 1–18. https://doi.org/10.1145/3586183.3606756

[85] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. 2022. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885* (2022).

[86] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The metacognitive demands and opportunities of generative AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–24.

[87] David H. Torres and Douglas E. Pruim. 2019. Scientific storytelling: A narrative strategy for scientific communicators. *Communication Teacher* 33, 2 (April 2019), 107–111. https://doi.org/10.1080/17404622.2017.1400679 Publisher: Routledge _eprint: https://doi.org/10.1080/17404622.2017.1400679.

[88] Barbara Tversky, Julie Bauer Morrison, Nancy Franklin, and David J Bryant. 1999. Three spaces of spatial cognition. *The Professional Geographer* 51, 4 (1999), 516–524.

[89] Gilson Luiz Volpato. 2015. O método lógico para redação científica. *Revista Eletrônica de Comunicação, Informação & Inovação em Saúde* 9, 1 (2015). https://doi.org/10.29397/reciis.v9i1.932

[90] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[91] Nedra Kline Weinreich. 2010. *Hands-on social marketing: a step-by-step guide to designing change for good*. Sage.

[92] Dustin J Welbourne and Will J Grant. 2016. Science communication on YouTube: Factors that affect channel and video popularity. *Public understanding of science* 25, 6 (2016), 706–718.

[93] Wikipedia contributors. 2024. Popular science — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Popular_science Accessed: 2025-03-27.

[94] 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.

[95] Haijun Xia, Hui Xin Ng, Zhutian Chen, and James Hollan. 2022. Millions and Billions of Views: Understanding Popular Science and Knowledge Communication on Video-Sharing Platforms. In *Proceedings of the Ninth ACM Conference on Learning@ Scale*. 163–174.

[96] Liwenhan Xie, Chengbo Zheng, Haijun Xia, Huamin Qu, and Chen Zhu-Tian. 2024. WaitGPT: Monitoring and Steering Conversational LLM Agent in Data Analysis with On-the-Fly Code Visualization. https://doi.org/10.1145/3654777.3676374 arXiv:2408.01703 [cs].

[97] Leni Yang, Xian Xu, XingYu Lan, Ziyan Liu, Shunan Guo, Yang Shi, Huamin Qu, and Nan Cao. 2021. A design space for applying the freytag's pyramid structure to data stories. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 922–932.

[98] Yu Zhang, Kexue Fu, and Zhicong Lu. 2025. RevTogether: Supporting Science Story Revision with Multiple AI Agents. *arXiv preprint arXiv:2503.01608* (2025).

[99] Yu Zhang, Changyang He, Huanchen Wang, and Zhicong Lu. 2023. Understanding Communication Strategies and Viewer Engagement with Science Knowledge Videos on Bilibili. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[100] Zheng Zhang, Jie Gao, Ranjodh Singh Dhaliwal, and Toby Jia-Jun Li. 2023. Visar: A human-ai argumentative writing assistant with visual programming and rapid draft prototyping. In *Proceedings of the 36th annual ACM symposium on user interface software and technology*. 1–30.

[101] Jelena Šuto, Ana Marušić, and Ivan Buljan. 2023. Linguistic analysis of plain language summaries and corresponding scientific summaries of Cochrane systematic reviews about oncology interventions. *Cancer Medicine* 12, 9 (2023), 10950–10960. https://doi.org/10.1002/cam4.5825

arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/cam4.5825

A Appendix

A.1 Specific Strategies for Science Communication Writing

Table 4. Design Space for Science Communication Writing

Category	Strategy	Definition	Label
Scientific Exposition	(1) Layered Transitions [51, 60, 75, 89]	Use multiple transition words or phrases (e.g., "but," "and," "therefore") within a short span to emphasize logical shifts and contrasts.	4
	(2) Rigorous Source Verification [2, 51, 73]	Cross-check scientific claims and data against reliable, peer-reviewed sources to ensure exposition.	3
	(3) Step-by-Step Explanation [3, 51]	Introduce the core idea first and then progressively add background details, creating a structured learning process.	2, 4
	(4) Acknowledge Uncertainties [69]	Transparently discuss uncertainties, potential biases, or limitations in data and models to build credibility.	1, 2
	(5) Consistent Terminology [52]	Use the same terminology throughout the content to maintain clarity and avoid confusion.	1
	(6) Citations & Quotes [2, 27]	Integrate citations and direct quotes seamlessly to enhance credibility while maintaining narrative flow.	3
	(7) Everyday Events to Scientific Insights [3, 52]	Automatically identify and link theories or knowledge to real-world events or stories mentioned in the text.	2, 3
Narrative Engagement	(8) Question-Answer Hook [29, 42, 53]	Ask a direct question and provide an immediate answer to introduce key concepts clearly and concisely.	5, 6, 7
	(9) Reflection Question [29]	Ask a thought-provoking question that does not require an immediate answer, encouraging reflection and reinforcing key concepts.	5, 7, 8
	(10) Suspense-Driven Reveal [95, 99]	Present a question, problem, or scenario at the beginning and delay its resolution to sustain curiosity.	5, 7
	(11) Use metaphors [25, 29, 52]	Convey unfamiliar concepts by drawing analogies to more familiar ones.	5, 6
	(12) Inject humor [39]	Use playful language or puns to make the content more engaging and enjoyable.	5, 8
	(13) Add real-world supporting examples [55, 57]	Illustrate abstract concepts using relatable, real-world examples.	5, 6
	(14) Add stories [17, 18, 57]	Use narratives with characters, settings, and plot progression to enhance engagement and memorability.	5, 6, 8
	(15) Add an imagery description [1, 29, 38]	Use vivid, sensory details to help the audience visualize concepts.	5, 6
	(16) Create negative emphasis for focused attention [29, 38, 42, 64]	Highlight extreme negative outcomes to intensify focus and reinforce key lessons.	5, 8
	(17) Make positive emotion to expand action repertoire [29, 33, 38, 64, 74, 91]	Use uplifting messages, particularly in conclusions, to inspire optimism and motivation.	5, 8
Both	(18) Simplify and abstract language [44, 48, 101]	Rephrase complex scientific terminology or detailed descriptions into more general, accessible language without compromising core exposition.	1, 6
	(19) Clarify Key Terms [64, 75]	Define complex or specialized terms at the beginning to establish a shared understanding.	1, 6
	(20) Key Point Recap [29, 64, 87]	Summarize the main points concisely at the conclusion of the content to reinforce memory retention.	1, 4, 6
	(21) Repeat key point(s) or question(s) [4, 47]	Reinforce key concepts by strategically repeating crucial terms or questions.	1, 6
	(22) Emphasize with Numbers [31, 97]	Connect scientific discussions to real-world recent news or trends to enhance relevance and engagement.	1, 2, 3, 8
	(23) Strengthen the Connections Between Content [60, 89]	Ensure smooth transitions between related ideas by using bridging statements or contextual links.	4, 6
	(24) Present Balanced Views [52]	Provide both supporting evidence and counterarguments to present a well-rounded discussion.	2, 6
	(25) Tie Science to Current Events [3, 52]	Connect scientific discussions to real-world recent news or relavant stories.	3, 5, 6

***Lable:** *Scientific Exposition Effects:* 1. Articulate Precisely; 2. Elaborate Thoroughly; 3. Verify Knowledge; 4. Maintain Logical Consistency

Narrative Engagement Effects: 5. Captivate & Immerse; 6. Enhance Understanding; 7. Inspire Curiosity; 8. Evoke Emotion

A.2 Rating Model Construction

Our primary goal in constructing the coordinate axis is to simulate audience feedback so that users can receive real-time evaluations. Therefore, we collected real user feedback on texts with varying characteristics to fine-tune a LLM that can provide scores during the real-time writing process.

Dataset Construction We first built a dataset of popular science texts containing 45 texts (example in section A.2.1) from five commonly seen science communication topics: psychology, economics, geography, history, and physics. For each topic, there are nine texts; three each of long (300 words), medium (150 words), and short (50 words) formats; representing three typical levels of revision granularity in science communication. Within each length category, we included three different levels of narrative transformation: (1) purely expository scientific texts (Expository), (2) fully narrative story-like texts (Story), and (3) an intermediate "infotainment" style (Medium), which is an ideal format in popular science that maintains scientific exposition while incorporating narrative strategies from our design space. All texts were revised by an expert with two years of experience in science communication writing

Score Collection We designed a survey to collect ratings for these texts on two dimensions: Narrative Engagement and Scientific Exposition, two main communication goals in popular science [16]. For Narrative Engagement, we used five subscales: Narrative Presence, Emotional Engagement, Narrative Understanding, Curiosity, and General Narrative Engagement, a survey developed by prior work [8]. For Scientific Exposition, given the lack of mature scales, we measured five dimensions inspired by standards for scientific texts from previous research [16]: Conceptual Clarity, Plausibility, Completeness, and Factual Correctness. When it comes to scientific exposition, our focus is more on the audience's subjective experience during reading rather than an objective verification of exposition. Since readers vary in their background knowledge, what we emphasize is not just factual correctness, but the perceived trustworthiness of how the content is presented — that is, how reliable and credible the text appears to them The full questionnaire can be found in the section .

Participants First, we recruited three experts (each with more than one year of experience in creating science narratives) to rate the texts. After rating, they discussed and jointly established a scoring rubric, including benchmarks for each score range from 0 to 10. Next, we recruited 27 participants interested in science communication. We invite experts to establish standards as a reference point for audience ratings, in order to reduce variance in their subjective evaluations of the text. The criteria established by experts are in the Appendix A.2.3.

Survey Results The distribution of scores for the 45 texts is displayed in the Figure 9. It is shown that story-like texts tend to elicit higher narrative engagement but exhibit lower scientific exposition. In contrast, expository texts maintain higher scientific exposition at the expense of engagement. The infotainment style appears to strike a balance between the two. Additionally, longer texts generally perform better in both dimensions, whereas shorter texts show lower overall scores, likely due to limitations in content depth and development.

Final Model Fine-Tuning For each text, we first computed the average score across the five questions within each of the two dimensions and then averaged these scores across all 27 participants. To match the 0–100 scale of the final coordinate axis, the scores were scaled by a factor of 10. These scaled scores (representing the two dimensions) served as the output, while the corresponding text and the expert-defined criteria used as reference formed the input.

During the development phase, we adopted a small-sample fine-tuning strategy to customize GPT-4o for our domain-specific application. This approach, which leverages a relatively limited number of high-quality training examples, has

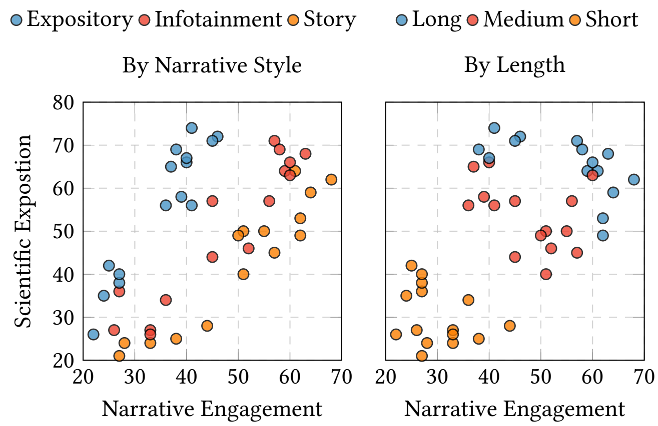

Fig. 9. Each point represents one of 45 science communication texts, plotted by its average audience rating for narrative engagement (x-axis) and scientific exposition (y-axis), based on 27 crowd-sourced rubric-based evaluations per text. The left panel groups texts by narrative style: Expository (informational, fact-focused), Story (highly narrative), and infotainment (represents infotainment-style revisions that blend factual exposition with narrative strategies). The right panel groups texts by length (Short=50 words, Medium=150 words, Long=300 words).

been shown to be both efficient and practically effective in enhancing model performance on specialized tasks [5]. We prepared and uploaded the curated dataset through OpenAI's official platform and used their fine-tuning API to tailor GPT-4o. The resulting customized model served as the backbone of our scoring system.

Technical Evaluation To validate the reliability of this scoring mechanism, we conducted a formal evaluation. We constructed a controlled dataset consisting of five source articles, each systematically rewritten into three different lengths (long, medium, short) and expressed in three different styles (expository, medium, story). This design yields nine distinct variants per article, resulting in a total of 45 text samples. From this dataset, we randomly selected 33 samples for fine-tuning GPT-4o, while reserving 12 samples for evaluation. The fine-tuned model was assessed against human ratings on two key dimensions: narrative engagement and scientific exposition. On the held-out test set, the fine-tuned model demonstrated a high degree of alignment with human judgment, achieving Pearson correlation coefficients of 0.90 and 0.91 for narrative and exposition scores, respectively. In addition, the model's predictive reliability was reflected in RMSE values of 6.48 and 7.02. These results indicate that the fine-tuned LLM scoring mechanism can effectively approximate human evaluative patterns, thereby providing a reliable and scalable alternative to manual scoring.

A.2.1 Example of Content.

Please view the materials via this anonymous link: https://cryptpad.fr/doc/#/2/doc/view/7V7gS5xcQdZwo0mLeBbfiQe6HEgU+02HqdaupBV9tA0/

A.2.2 Survey used for gathering audience feedback.

Please view the survey via the anonymous link: https://cryptpad.fr/doc/#/2/doc/view/XfWs-wD3qmBXSnEC0YqM9EZg2GO++H2RJYUqyrcvj1I/

[5]https://platform.openai.com/docs/guides/fine-tuning?utm_source=chatgpt.com

A.2.3 Score Criteria.

Please view the criteria via this anonymous link: https://cryptpad.fr/doc/#/2/doc/view/uNMusLpCPWGwzqKWi04F0TY+20nW2hnG1NkS1V2BHB4/

A.3 Materials used for experiment

Please view the materials via this anonymous link: https://cryptpad.fr/doc/#/2/doc/view/Q3Jhj+HhzHtt9zYqyF0Sv4mziQYBp6oWl43a84Gqmeg/

A.4 Survey

Part 1: Metacognition

Metacognitive Knowledge: This pertains to an individual's awareness and understanding of their own cognitive processes and strategies

Q1: I am aware of my writing goals during the editing process.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Metacognitive Regulation: This involves the active management of one's cognitive processes through planning, monitoring, and evaluating

Q2: I set specific goals for what I wanted the narrative to achieve.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q3: I reflect on my writing strategies or editing choices while using the AI writing tool. (Indicates real-time assessment of strategy effectiveness.)

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q4: During writing, I regularly checked whether the narrative was staying on track with my intended message.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q5: I can clearly identify areas of my writing that need improvement when using the AI tool.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q6: After writing, I reviewed the narrative to assess how well it communicated the scientific content.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q7: I am able to adjust my writing strategies during the editing process.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Part 2: Control (Control:)

Q8: I felt in control of the writing process while interacting with the system.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q9: I was able to override or ignore the system's suggestions when I thought it was necessary.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q10: I determined the direction and flow of the science narrative, not the system.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Part 3: Autonomy (Autonomy:)

Q11: I felt free to make my own choices during the co-writing process with the system.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q12: The system supported my ability to express my own ideas in the narrative.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q13: I did not feel pressured to accept the system's suggestions.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

A.5 Participants demographic information

.

ID	Age	Gender	Education	Science Communication	AI Writing Use	Writing Confidence	Occupation
1	26	Male	Postgraduate	Experienced Creators	Occasionally	Confident	(a)
2	27	Male	Postgraduate	Expert	Daily	Confident	(a), (b), (c), (d)
3	26	Male	Postgraduate	Experienced Creators	Daily	Confident	(b), (d)
4	25	Female	Postgraduate	Experienced Creators	Daily	Confident	(a), (b), (c)
5	24	Male	Postgraduate	Experienced Creators	Daily	Confident	(a)
6	28	Female	Postgraduate	Senior Audience	Weekly	Neutral	(a)
8	28	Male	Postgraduate	Senior Audience	Occasionally	Neutral	(a)
7	29	Female	Higher than postgraduate	Experienced Creators	Daily	Confident	(a), (b)
9	31	Male	Postgraduate	Experienced Creators	Weekly	Neutral	(a)
10	24	Female	Postgraduate	Experienced Creators	Occasionally	Confident	(a), (c)
11	29	Female	Postgraduate	Experienced Creators	Weekly	Neutral	(a)
12	26	Male	Postgraduate	Experienced Creators	Weekly	Neutral	(a)
14	27	Male	Postgraduate	Experienced Creators	Daily	confident	(a), (b)
15	24	Female	Postgraduate	Senior Audience	Weekly	Neutral	(a)
16	30	Male	Postgraduate	Experienced Creators	Weekly	Neutral	(a)

Occupation: **(a) PhD Student / Postdoctoral Researcher/University Faculty / Researcher;**
(b) Science Journalist / Media Producer;
(c) Educator / Teacher;
(d) Online science Content Creator (e.g., YouTube, Blog, TikTok, etc.)

A.6 User Study Results

1. Visualization of interaction behaviors from 16 participants across two revision directions:

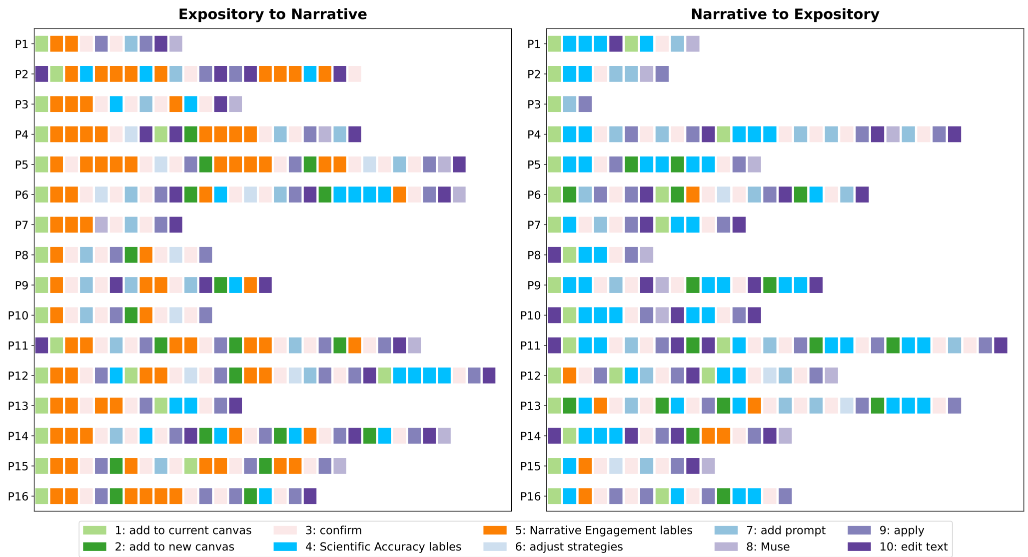

Fig. 10. Visualization of interaction behaviors from 16 participants across two revision directions.

2. Functional of SpatialBalancing evaluation results:

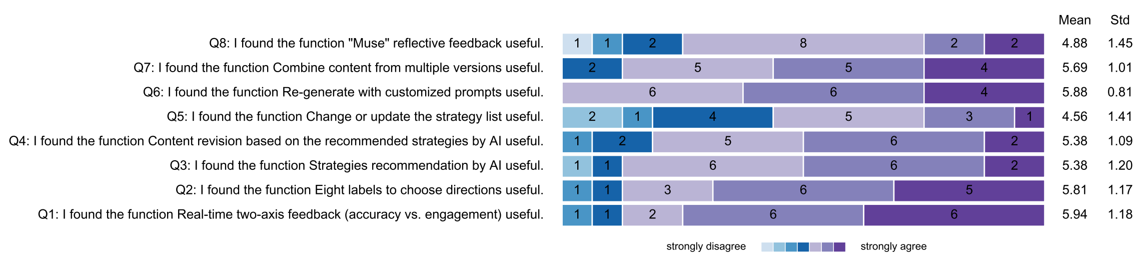

Fig. 11. Functional Evaluation of SpatialBalancing.

A.7 Prompts

A.7.1 Recommender.

The blue word will be replaced by input information.

```
# Base prompt
You are an expert in science communication narrative text revision and strategy recommendation.
Your task is to analyze the given text and recommend effective strategies to improve it.

# Order prompt
Step 1: Analyze the Text.
Position: Identify where the selected text {text} appears in the {overall_content}.
Granularity: Determine whether the text consists of sentences, paragraphs, or a complete document.
Core Message: Extract the key ideas that must be preserved and effectively conveyed in text.

Step 2: Select Strategies Review the available strategy list {strategy_info}, including their
definitions, examples, and usage instructions. Choose a set of strategies that align with the
text's characteristics and modification goals. Ensure the selected strategies are compatible
when combined. Consider multiple ways to apply the strategies for improvement.
Only choose strategies mentioned above, and use them appropriately.
Provide {generated_number} different versions, each using distinct or complementary strategy sets.
These different versions should use different strategies, preferably with varied combinations of
strategies.

Step 3: Output the Strategy List Return the strategy selection in JSON format with multiple versions:
{
"Version1": [ "Strategy_A", "Strategy_H", "Strategy_J", "Strategy_B"],
"Version2": [ "Strategy_F",..., "Strategy_E"],
...,
"Version_number": [ "Strategy_G", "Strategy_M",..., "Strategy_C",...,"Strategy_D"]
}
Do not include any extra commentary or explanation outside the JSON.
Let's think step by step.
```

A.7.2 *Generator.*

The blue word will be replaced by input information.

Generate new text based on user selected goals

```
# Order prompt
You are an expert in science communication narrative strategy. Your task is to revise the
given text using the recommended strategies and provide a concise overview of how the
strategies were applied.

Step 1: Review the Strategy List
- Read the strategy list {strategy_info}, including each strategy's definition and
how it is typically used.

Step 2: Apply all the Strategies mentioned in the strategy list to the Text: {text}.
Even if the original text already contains elements that align with the strategy, enhance it further
based on how the strategy should be applied.
Also, consider the position of the given text in the whole context {overall_content}.
Make the changed text coherent with the context.

Step 3: Summarize the Application
- Summarize how each selected strategy was applied.
- Keep the summary concise and short to indicate what specific changes have been made using
separate strategies.

Step 4: Do not omit or alter any important information from the original text, but ensure that the
generated text is distinct from the original.

Step 5: If the content is primarily narrative in nature, supplement it with scientifically grounded
explanations, relevant data, or reliable sources to enhance credibility and depth.

Step 6: Output the Result Return a JSON with the following structure:
{
"strategies": ["Strategy_A", ..., "Strategy_B", "Strategy_C", "Strategy_D"],
"summary": "Summarize how each strategy was applied and what specific changes were made to the content
             based on each strategy. Example: Changed 'Photosynthesis is the process plants use to
             make food.' to 'What if plants could teach us how to turn sunlight into fuel?
             Focus only on the changes from the previous version.'",
"newText": "Modified version of the text. Even if the original text already contains elements that
             align with the strategy, enhance it further based on how the strategy should be applied."
}

Do not include any extra commentary or explanation outside the JSON.
Let's think step and step.
```

A.7.3 Scorer.

The blue word will be replaced by input information.

```
# Base prompt
You are an engaging audience for science communication.
Given a narrative, evaluate it on two dimensions: (1) Narrative Engagement and (2) Scientific Exposition.
using the detailed scoring rubrics below.
Provide a numerical score from 0 to 100 for each dimension, along with a brief explanation justifying
your rating.

Dimension 1:
Narrative Engagement: Evaluate how effectively the narrative captures attention, evokes emotion,
sparks curiosity, and maintains reader engagement.
Scoring Rubric:
0-20: Extremely boring and dry, no storytelling elements,
21-40: Barely engaging, logical but lacks emotion or creativity,
41-60: Moderately engaging, uses some analogies or description but still feels academic,
61-80: Quite engaging, includes storytelling techniques and relatable examples,
81-100: Highly immersive, vivid storytelling with strong emotional or narrative appeal.


Dimension 2: Scientific Exposition: Assess how well the narrative explains scientific concepts with
clarity,
correctness, and alignment with established knowledge.
Scoring Rubric:
0-20: Highly inaccurate or pseudoscientific, major factual errors,
21-40: Misleading or speculative, lacks clarity or evidence,
41-60: Mostly accurate but vague or oversimplified,
61-80: Generally accurate, minor imprecision, lacks citations,
81-100: Highly accurate, precise, and well-aligned with scientific consensus.

# Order prompt
This is the original text: {text} and its score {currentScore}. Please use this as a reference.
Compare the current version with the original one in terms of scientific exposition and narrative
engagement, and assess whether it performs better or worse than the previous version.
Compared to the previous version's scores, assign a score difference within a reasonable range.
```