Compositional Structures as Substrates for Human-AI Co-creation Environment: A Design Approach and A Case Study

Yining Cao
University of California, San Diego
San Diego, California, USA
yic069@ucsd.edu

Yiyi Huang
University of California, San Diego
San Diego, California, USA
yh89360@gmail.com

Anh Truong
Adobe Research
New York, California, USA
truong@adobe.com

Hijung Valentina Shin
Adobe Research
Cambridge, Massachusetts, USA
vshin@adobe.com

Haijun Xia
University of California, San Diego
San Diego, California, USA
haijunxia@ucsd.edu

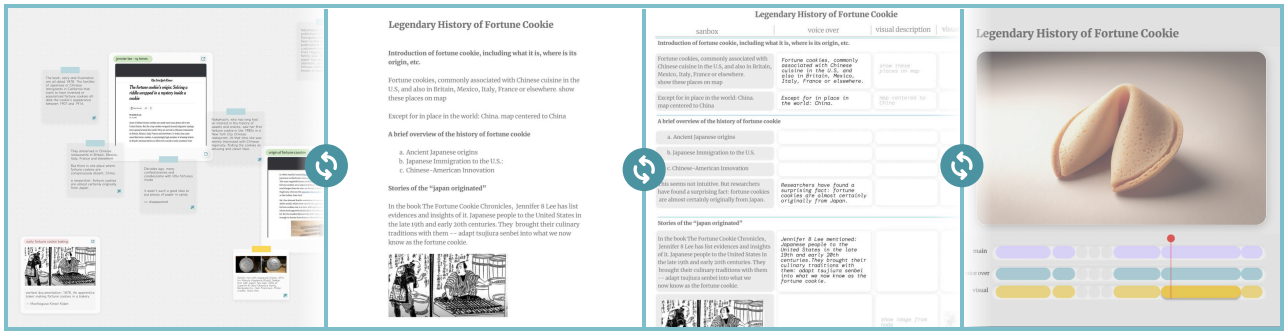


Figure 1: The user interface of VideOrigami, a human-AI video co-creation environment, developed using the proposed approach of combining compositional structures and AI. The compositional structures facilitate inspection and control of AI generation, and AI facilitates information transformation and synchronization within and across the structures.

Abstract

It has been increasingly recognized that effective human-AI co-creation requires more than prompts and results, but an environment with empowering structures that facilitate exploration, planning, iteration, as well as control and inspection of AI generation. Yet, a concrete design approach to such an environment has not been established. Our literature analysis highlights that compositional structures—which organize and visualize individual elements into meaningful wholes—are highly effective in granting creators control over the essential aspects of their content. However, efficiently aggregating and connecting these structures to support the full creation process remains challenging. We, therefore, propose a design approach of leveraging compositional structures as the substrates and infusing AI within and across these structures to enable a controlled and fluid creation process. We evaluate this approach through a case study of developing a video co-creation environment using this approach. User evaluation shows that such an environment allowed users to stay oriented in their creation activity, remain aware and in control of AI's generation, and enable flexible human-AI collaborative workflows.

CCS Concepts

• **Human-centered computing** → **Graphical user interfaces**; **Natural language interfaces**; **Collaborative interaction**.

Keywords

Design Approach, Compositional Structures, Human-AI Collaboration, Video Creation

ACM Reference Format:
Yining Cao, Yiyi Huang, Anh Truong, Hijung Valentina Shin, and Haijun Xia. 2025. Compositional Structures as Substrates for Human-AI Co-creation Environment: A Design Approach and A Case Study. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3706598.3713401

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713401

1 Introduction

Content creation is inherently iterative, involving exploration, planning, and refining. While advanced AI models are capable of generating high-quality text, images, and video clips from prompts [2, 14], the HCI community has argued that relying on the prompt-generation paradigm alone is inefficient due to the lack of controllability and interpretability desired in the creative processes [108, 109]. Prior works have explored leveraging various external structures to augment human-AI collaboration: such as leveraging a chain structure to break down a complex task into smaller steps [100], employing content-specific structures to control AI generation (e.g., narrative structure in writing [50, 111]), and organizing generated text using diagrams and hierarchical structures for comprehension [48, 78].

These prior works collectively suggest that effective human-AI co-creation of complex content (e.g., scientific writing, music, narrative video) goes beyond simple prompt-generation cycles, requiring an environment with empowering structures to ground AI generation, facilitate human ideation, and support design iteration [62, 71, 76, 100]. Yet, there has not been an explicitly formulated approach to guide the development of such environments. Specifically, we lack systematic guidance on (1) what is the design process to follow, (2) what are the essential structures to consider, (3) how should AI be integrated with these structures, and (4) what are the benefits and challenges of such an environment. This work attempts to answer these questions.

Toward this goal, we surveyed prior research that investigated challenges and developed systems to support creative activities in a variety of domains, including writing, multimedia posts, podcasts, music, and video production. From this analysis, we identified a common approach in designing interfaces for supporting various forms of content creation: the use of *compositional structures*. We refer to compositional structures as structures that visualize and organize *individual components* of content into a cohesive and meaningful whole based on specific *content aspects*.

Our analysis revealed that compositional structures address four key content aspects: spatial (e.g., layout in graphical design), temporal (e.g., pacing in videos), narrative (e.g., storytelling coherence), and congruent (e.g., integration of multimodal elements such as text, visuals, and audio). For example, a narrative graph represents storylines as nodes and edges, enabling creators to inspect the flow of narrative points and experiment with alternatives; a multi-track timeline organizes individual video and audio clips along a temporal axis, supporting creators in sequencing, aligning, and adjusting the pacing of the clips. These structures not only assist in organizing and editing content by defining individual components and their organizational rules, but also provide functional affordances that guide creators through complex workflows, enabling efficient inspection, iteration, and refinement. While these structures can be employed individually, complex creative processes often require multiple structures to interoperate. For example, in narrative video creation, a timeline may synchronize with a storyboard to ensure the alignment between visual sequences and story progression.

Informed by the literature analysis, we propose a design approach for developing human-AI co-creation environments, which consists of four steps (1) identifying relevant compositional structures and their desired interconnections, (2) designing individual structures tailored to content aspects and workflow requirements, (3) aggregating these structures into a unified environment, and (4) infusing AI to support content creation and synchronization. With this approach, we aim to provide actionable guidance for building environments that balance human agency with AI augmentation, enabling effective human-AI co-creation.

An ideal evaluation of a design approach is to test it across multiple domains. This is challenging in terms of scope, as building an environment to support a single domain's extended workflow demands substantial design and development effort. Therefore, we opted for a case study in video creation. This domain encompasses the four content aspects identified in our literature analysis—spatial, temporal, narrative, and congruent—and thus warrants reasonable generalizability. We conducted a formative study to identify common compositional structures in video creation workflows as well as practices and challenges associated with these structures. We then developed a human-AI video co-creation environment, VideoOrigami, by infusing these compositional structures with AI. We evaluated this co-creation environment by conducting a user evaluation with ten video creators. This study enabled us to investigate the benefits and challenges of such an environment and discover new creation patterns resulting from the reduced cost of aggregating compositional structures and integrating AI.

Together, this work makes the following contributions.

- **A literature analysis** of prior work across multiple creative domains, identifying compositional structures as a foundational design element for human-AI co-creation environments and summarizing the design practices, challenges and opportunities of leveraging compositional structures.
- **A design approach** that proposes using the compositional structures as substrate for human-AI co-creation environment to ground the generation process; and infusing AI within and across these structures to enable flexible creation workflows.
- **A case study** where we developed a human-AI video co-creation environment, demonstrating the feasibility of instantiating the proposed approach.
- **A user evaluation** of the developed environment, validating the effectiveness of the approach and uncovering new patterns of human-AI collaboration enabled by the integration of compositional structures and AI.

2 Related Work

Our research proposes a human-AI collaboration paradigm by grounding AI automation with compositional structures. We herein review prior work on human-AI collaboration and the role of compositional structures in scaffolding content creation. Additionally, we examine relevant literature on video creation.

2.1 Supporting Human-AI Collaboration

Recent advancements in generative AI are shifting content creation from relying on low-level manual editing to guiding AI with high-level instructions and goals [3, 16, 81]. This is a significant step towards the human-AI symbiotic collaboration that Licklider envisioned [56]. To support effective human-AI collaboration, recent

work has made progress on various fronts, including design guidelines [7, 45], analytical frameworks [74, 82], and interaction techniques [48, 98]. For example, Amershi et al. proposed guidelines for designing human-AI interaction, which describe the desired high-level qualities of human-AI interfaces, such as "show contextually relevant information" and "learn from user behavior" [7]. Terry et al. and Subramonyam et al. proposed extending Norman's Gulfs of Execution and Evaluation with Process Gulf [82] and Envisioning Gulf [74], respectively, to better model human-AI interaction.

Recognizing the lack of control and interpretability of the prompt-generation paradigm, prior work explored incorporating various structures into human-AI interaction, such as breaking down complex tasks into granular steps [100], visualizing the generation space using key content dimensions [15, 76], adding additional visual structures (e.g, human poses) to control image generation [110], using narrative structure [111], node-link structures (e.g.,diagrams [48], and hierarchical spatial structures [78]) to organize generated text.

We extend the prior work on employing structures to improve human-AI interaction. Beyond utilizing individual structures for specific tasks, we explore how to develop co-creation environments enriched with these structures [53, 62, 71]. Buschek's work on AI writing tools offers a related perspective, identifying an interface design pattern of collaging fragmented views to create the interfaces of writing systems [17]. Our approach also draws on the concept of information environments built from "information substrates" [11, 34, 54, 107]. We propose employing compositional structures as a type of substrates, and infusing AI within and across them to create the human-AI co-creation environments.

2.2 Effectiveness of Compositional Structures

Compositional structures that describe the form, arrangement, and relationships of components have been found to be effective in facilitating the creation, consumption, evaluation, and iteration of information content [18, 52, 55, 69].

Visualizing compositional structures can help users develop an overall understanding of the content and enable efficient consumption. For example, exploded-view drawings are effective in communicating the composition of various parts to inform the assembly, disassembly, and repair of mechanical components [55]. The table of contents in books enables readers to quickly review the structure of a book and facilitate navigation. Similar structures have been adopted to assist in the consumption of videos by segmenting long-form videos into smaller and skimmable chunks based on narrative structures, such as SceneSkim [64] and Video digests [65].

During content creation, compositional structures enable creators to inspect and define the structures of the content and support structural revision [33, 50]. For example, exploring the composition of ideas during pre-writing was found effective in improving writing quality [33], and reverse outlining can effectively aid in the structural revision of the documents [52]. For music, compositional structures have been devised to facilitate the composition of chords [37] and the entire music score [38]. For filmmaking, the three- or five-act narrative structures are established practices to create compelling narratives [32]. Inquiry-based structure is commonly used in science communication videos to maintain viewer engagement [105]. Screenplay, storyboards, and audio-video scripts are commonly used to plan and develop films and videos [32, 58].

Given the effectiveness of compositional structures, recent work has sought to leverage them to assist human-AI co-creation. For example, research explored leveraging compositional structures to help writers plan, organize, and revise their writing with AI, such as using AI to generate outlines [26] or to generate passages based on argumentative structures [111]. Metaphorian supports creating extended metaphors by allowing users to define the structures of concepts they want to explain and leverage AI to generate sets of concepts exhibiting congruent structures [50]. These works, however, typically focus on utilizing a single compositional structure. In contrast, most content creation involves extended workflows that interleave many compositional structures in a highly dynamic and contingent manner [61, 75]. Therefore, we explore how to develop human-AI co-creation environments with multiple compositional structures that support an entire creation workflow.

2.3 Supporting Video Creation

Video is a highly versatile medium that can integrate various forms of content, such as images, animations, text, infographics, sketches, sounds, and recordings. Composing these numerous heterogeneous materials of different modalities and formats into a coherent audio-visual piece is a highly challenging and tedious task, and therefore, received significant attention from HCI.

Researchers and practitioners have proposed principles and practices regarding video composition, and developed many authoring support systems based on them. For example, the Congruence Principle states that the content and format of the visual content should be congruent to those indicated in the narrative [87]. Leveraging the desired congruence between the visual content and the underlying narrative, research has explored offloading the tedious, frame-level interactions of clips such as visual search [46, 103], cut placement [12, 84], and clip sequencing [67, 68, 91, 95] with the synchronized manipulation of the corresponding scripts [12, 101]. For example, Quickcut enables the automatic assembly of shots into a whole video by temporally aligning the shots with the video script [84]. Crosspower leverages the semantic structures in the scripts to facilitate the spatial composition of visual materials in the scene [101]. Research has also explored leveraging the compositional structures of existing content in other media to generate videos. For example, end-to-end systems have been developed to create videos by transforming the composition of static content (e.g., documents, webpages), such as generating TikTok videos from news articles [96] and marketing videos from websites [23].

We developed a human-AI video co-creation environment following our proposed design approach. As we will demonstrate, the interconnected and intelligent compositional structures serve as a versatile interface foundation that can coherently support many of the existing techniques but also afford new ones in the context of human-AI co-creation. Additionally, our user evaluations provide new insights into the challenges and opportunities in human-AI co-creation. The notorious complexity of video content and its workflow make video creation an ideal domain for validating the design approach, warranting considerable generalizability of the proposed approach and the resulting findings.

3 Compositional Structures as Substrates: A Design Approach

We grounded the development of the design approach in the existing literature dedicated to designing interactive systems for content-creation activities. Specifically, we sought to understand: *how existing systems design compositional structures to support the development of the various aspects of the content.* We analyzed work targeting diverse domains, including writing, music, podcasts, interactive media, and video. These domains allowed us to comprehensively cover textual, visual, audio, and interactive content. For each domain, we selected survey, study, and system articles as seed articles and utilized a snowball method to collect relevant literature.

For each article, we annotated the compositional structures that were studied, along with the content aspects they aim to support and their functionality. We extracted excerpts from the papers regarding the design decisions and challenges associated with the structures. We stopped the snowball process when no new compositional structures were found. In cases where multiple articles addressed the same compositional structure, we prioritized those that were widely cited. Ultimately, we identified 55 papers, and more details of these papers are included in the Appendix A.3.

3.1 Definition of Compositional Structures

Our literature analysis revealed that interfaces for supporting content creation are typically designed with one or more structures. These structures define the individual components and organizational rules for assembling the components based on specific aspects of the content, offering corresponding functional affordances that facilitate the composition process – we refer to these structures as **compositional structures**. By analogy with biological substrates—surfaces on which living organisms grow—we conceptualize compositional structures as substrates on which informational content 'grows'.

We herein summarize the **utility of compositional structures** across key aspects of content creation, analyzing the design of their individual components, organizational rules, and functional affordances implemented in existing systems (Section 3.2). Among the reviewed papers, 47 out of 55 combined multiple compositional structures to support the creation process. By analyzing the **interleaving usage of different compositional structures**, we summarize current design practices for establishing synchronization of these structures and formalize the solutions for aggregating them into one workspace. Additionally, we highlight challenges identified in existing research and propose opportunities to leverage AI to support the iterative creation processes across multiple compositional structures (Section 3.3). This analysis provides insights into *what* compositional structures should be integrated into co-creation environments and *how* to infuse AI within and across structures to support creative workflows.

3.2 Utility of Compositional Structures

Each compositional structure assists in the creation of one or more aspects of the content. We categorized four key aspects: spatial, temporal, narrative, and congruent.

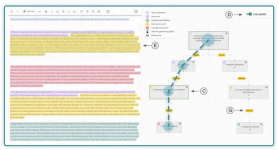
3.2.1 Spatial Aspect. This aspect primarily refers to the layout of individual components in the final content. It pertains to the organization of different elements in visual content for effective communication and appeal. Compositional structures supporting this aspect are often based on a *free-form canvas* or *grid system* with directly manipulable elements, allowing flexible adjustments of their position and size. The organizational rules in these structures aim to satisfy constraints and preferences specific to the creation context, such as design guidelines, stylistic choices, or screen sizes. The compositional structures reify these rules to enable easy reuse and adaptation by both creators and automated processes, including auto-suggested templates or computational layout adaptations.

3.2.2 Temporal Aspect. This aspect addresses the pacing of content that unfolds over time. Interfaces designed to support this aspect typically feature a *multi-track timeline* that allows creators to place elements along a time axis, define keyframes, and preview the progression. While pacing is critical for effective storytelling, most systems require the creators to manually achieve the ideal pacing, such as deciding the duration of shots and adding pauses to enhance narrative impact. Prior work has explored implementing organizational rules based on desired dialogue styles on pacing [90]. Other works explored supporting element alignment by time, such as synchronizing video segments with transcripts [19, 46, 84]. Such synchronization embedded in the timeline accelerates the creation process by alleviating the manual coordination efforts otherwise required to deal with different tracks in the timeline structure.

3.2.3 Narrative Aspect. This aspect refers to the underlying story conveyed by the content. Creators often need to experiment with different narratives to develop the most effective one. Compositional structures supporting this aspect vary in levels of abstractions: from conceptual elements like the luckiness of a character in storytelling [24] to argumentative outlines in scientific writing [111]. The organization rules and functional affordances of these structures depend on both the content types and the creative workflow. For example, *graph structures* on a freeform canvas are commonly used for ideation, with association functions to generate or connect related concepts for brainstorming. *Hierarchical linear structures* are effective for planning and refinement, with summarization and expansion functions, such as summarizing paragraphs into a reverse outline or expanding simple headings into narrative points.

3.2.4 Congruent Aspect. This aspect addresses the integration of multiple individual elements perceived simultaneously during content consumption, such as notes in a musical chord [37], different modalities in a video [58, 84], and data visualizations and accompanying narratives in data storytelling [18]. To support this, compositional structures often associate these elements within a container to support reasoning: *lines* connecting notes in a musical chord [37], *storyboard cards* associating visuals with text [58], and *blocks* combining visualizations and descriptions side-by-side [18]. These containers frequently incorporate auto-update functionalities to ensure synchronized updates among elements and may include functionalities for suggesting alternatives or auto-completing based on congruence rules. Furthermore, by bundling elements,

Table 1: Compositional Structures for Four Content Aspects with Example Systems from Literature Analysis

Content Aspect	**Spatial Aspect** Spatial layout of individual elements within a content	**Temporal Aspect** Pacing of content that unfolds over time	**Narrative Aspect** Underlying message or story conveyed by the content	**Congruent Aspect** Integration of elements perceived simultaneously
Content	Multi-Media Document	Narrated Video	Argumentative Writing	Animated Data Story
Example System	**Adaptive Layout** [Schrier et al. 2008] 	**Quickcut** [Truong et al. 2016] 	**Visar** [Zhang et al. 2023] 	**DataParticles** [Cao et al. 2023] 
Compositional Structure	Canvas-Based Layout Editor	Time-Aligned Transcript	Node-Link Graph for Visual Argumentative Outline	Block-Based Editor
Individual Component	Rectangular regions arranged on the page and filled with content	Raw footage segments with voice annotation	Nodes representing 5 types of argument such as main argument, counter argument, etc.	Text cell for story narrative; Visualization cell for animated unit visualization
Organization Rule	Layout of the elements adheres to template-based constraints and preconditions	Raw segments should be temporally aligned with the narration of the entire video based on semantic relevance	Nodes are connected with a set of logical relationships and organized hierarchically	Within a block, the visualization cell's encoding and animation should conform to description in the narrative text cell
Functional Affordance	Aggregate multi-source documents with adaptive layout Customize the layout on the fly	Match story narration with relevant video segments Choose frame-level video cut points adhering to film editing guidelines	Configure logical/hierarchical relationship Generate suggested arguments Update dependent arguments recursively	Generate animated unit visualization with narrative Customize the encoding and animation effects

creators can manipulate the container as a whole without compromising its internal congruence. Therefore, such structures often integrate congruence with other aspects, like narrative: for instance, a two-column structure, commonly used in video creation, expresses congruence within individual rows while supporting narrative progression through the linear arrangement of rows in a table format.

Notably, a single compositional structure can support multiple aspects simultaneously. For example, a transcript-based timeline supports both the temporal and narrative aspects as they have strong correspondence in certain video content. Similarly, the congruent aspect is often intertwined with others, as it inherently involves relationships between elements of different modalities. It is also important to acknowledge that while the spatial, temporal, narrative, and congruent aspects are prominent in the literature we have reviewed, they do not represent an exhaustive set of content aspects. For example, emotional resonance or interactivity may constitute additional dimensions that require further exploration.

3.3 Interleaving Usage of Compositional Structures and Challenges

Content creation is often multifaceted and highly iterative [18, 20]. These structures, however, are typically distributed across separate applications, resulting in fragmented workflows [18, 21, 40]. From the papers that utilized multiple compositional structures, we summarized two primary needs: *inspecting different content aspects* and *facilitating iterative transitions*. To address these needs and mitigate the challenges of fragmented workflows, a common design approach is to consolidate essential compositional structures into a single interface, and establish synchronization among them. To inform our design approach, we summarize existing practices for integrating compositional structures and achieving synchronization across these structures.

As each compositional structure defines its individual components and their organization rules, the synchronization aims to establish connections between the components across structures while satisfying their respective rules. We summarized three aspects to consider for establishing synchronization and highlighted the design challenges within each aspect.

3.3.1 Defining Correspondences Between Individual Units Across Structures. Establishing how individual components in one structure correspond to those in another is fundamental to achieving synchronization across structures. For example, a node in the narrative graph corresponds to a sentence in the text editor [24, 111]. Predefined correspondences help users understand the system's logic and how modifications in one structure influence others. However, they can also limit flexibility. As the creative process evolves, these correspondences may become ambiguous, requiring reevaluation or reconfiguration, which can disrupt the creative flow.

3.3.2 Determining Appropriate Synchronization Techniques. We identified two common techniques for cross-structure synchronization: *Synchronized Highlighting*, which highlights corresponding units across structures to aid navigation and reference, and *Synchronized Editing*, where updates in one structure propagate to corresponding units in another. Systems with synchronized editing typically also include synchronized highlighting. The complexity of synchronized editing depends on how content is represented across structures. It may involve direct content transformation (e.g., transferring a phrase in an outline to a section heading), attribute-based transformation (e.g., a text snippet converted to a timeline

duration), or advanced AI-driven generation (e.g., a paragraph converted to an outline point). While advanced transformations can streamline workflows, they often require significant review and adjustments. Simple synchronized highlighting may be preferable when creators prioritize manual control over automation.

3.3.3 Configuring Updating Mechanism. Two key factors influence the updating mechanism: (1) Directionality: updates can be *bi-directional or one-directional*, where changes flow between structures in both directions or only from one to another; (2) Control: whether updates should be *automated or user-controlled*, with automated updates ensuring consistency but potentially causing unintended overwrites, while user-controlled updates provide greater control but require additional actions that might be laborious. While most systems favor automated bi-directional updates for ease of use, this approach can pose challenges in preserving intentional modifications of content within one structure.

3.4 Proposed Design Approach

Existing research demonstrates the utility of compositional structures and explores practices for constructing workspaces that leverage them. Insights from this literature inform key considerations for designing such structures and integrating AI functionalities within and across them. Based on these findings, we propose the following four-step design approach:

- **ST1 Identifying Compositional Structures and Their Desired Interconnections of a Creation Activity.** The first step is to investigate the creation workflow for the targeted content. This includes identifying specific compositional structures used to address different content aspects, their roles at various workflow stages (e.g., ideation, editing, integration), and how these structures should interconnect to support transitions between them.
- **ST2 Designing Individual Structures with Content Aspects and Workflow Requirements.** For each identified structure, specify the individual components creators will manipulate and define the organizational rules (e.g., hierarchical relationships, temporal sequencing). The design of these structures should align with the intended creative outcomes and support efficient manipulation and arrangement within the structures.
- **ST3 Aggregating the Compositional Structures as the Foundation for the Co-creation Environment.** The defined structures need to be integrated into a workspace with desired synchronization by defining corresponding units, synchronization techniques (e.g., synchronized highlighting or editing), and updating mechanisms (e.g., one or bi-directional).
- **ST4 Infusing AI Functionalities within and across Compositional Structures to Facilitate Content Creation and Synchronization.** Based on the challenges within the creation workflow, automation needs to be infused within and across the structures. Within structures, AI should facilitate creating individual units adhering to their organizational rules; across structures, AI should maintain context awareness by managing references, coordinating interconnected content, and ensuring synchronization through appropriate updating mechanisms.

4 Understanding the Use of Compositional Structures in Video Creation

We evaluate our design approach by executing the proposed steps. We begin with identifying the compositional structures and the desired interconnections. As indicated in the previous section, many prior works have been conducted to understand and alleviate the challenges in video editing [22, 84, 103]. However, there lacks work that specifically investigates compositional structures in the entire video creation workflows, especially the desired interconnections among them. Therefore, we conducted an interview study with video creators to holistically understand their creation processes.

4.1 Participants and Procedure

We interviewed five expert video creators, each with over five years of experience in video production and publishing. To ground the interviews in concrete examples, we asked the creators to share videos they had produced, along with any materials used for planning and prototyping. Participants were purposefully selected to ensure their videos covered diverse types of content. In total, the interviews covered 14 videos: 3 vlogs, 3 short films, 3 explainers, 3 video essays, and 2 animated films. The interviews began with questions regarding creators' professional experiences, followed by an in-depth discussion of the creation processes. Conducted via video calls, each interview lasted around 90 minutes.

4.2 Compositional Structures in Existing Workflow

Despite the diverse video types, creators' workflows revolve around four key compositional structures shown in Fig. 2b. We first review these structures and then dive into the challenges of working with these structures.

4.2.1 Ideation and Asset Organization with Freeform Canvas. "You could see there is no discipline here, because I am just throwing them all in." (E2). All creators use dedicated spaces to organize relevant assets, such as documents, videos, and images. As creators sift through the assets, they jot down associated notes, develop narrative points, and connect them to form the storylines. Freeform canvases, such as Milanote (E1) and Miro board (E3, E5), were useful as they not only serve as the asset repository but also an ideation space where creators can arrange all materials to develop an understanding of the story. Creators frequently revisit this space to re-contextualize themselves with the materials.

4.2.2 Narrative Development with Linear Text Editor. "I just write out something almost off the top of my head with an idea." (E2) Creators need a space to organize disconnected ideas into a cohesive storyline. At this stage, a linear structure can be helpful, as it "actually helps me thinking." (E1) This stage involves substantial iteration and engagement with content of mixed fidelity. Creators usually begin by putting down ideas and talking points that emerge during ideation and asset collection. The initial content could be "written partially in full text, partially in outlining text." (E3) and as "a combination of scripts and visuals"(E2, E3, E5). Creators navigate through the content of mixed fidelity and modality and iteratively mold it toward a final storyline.

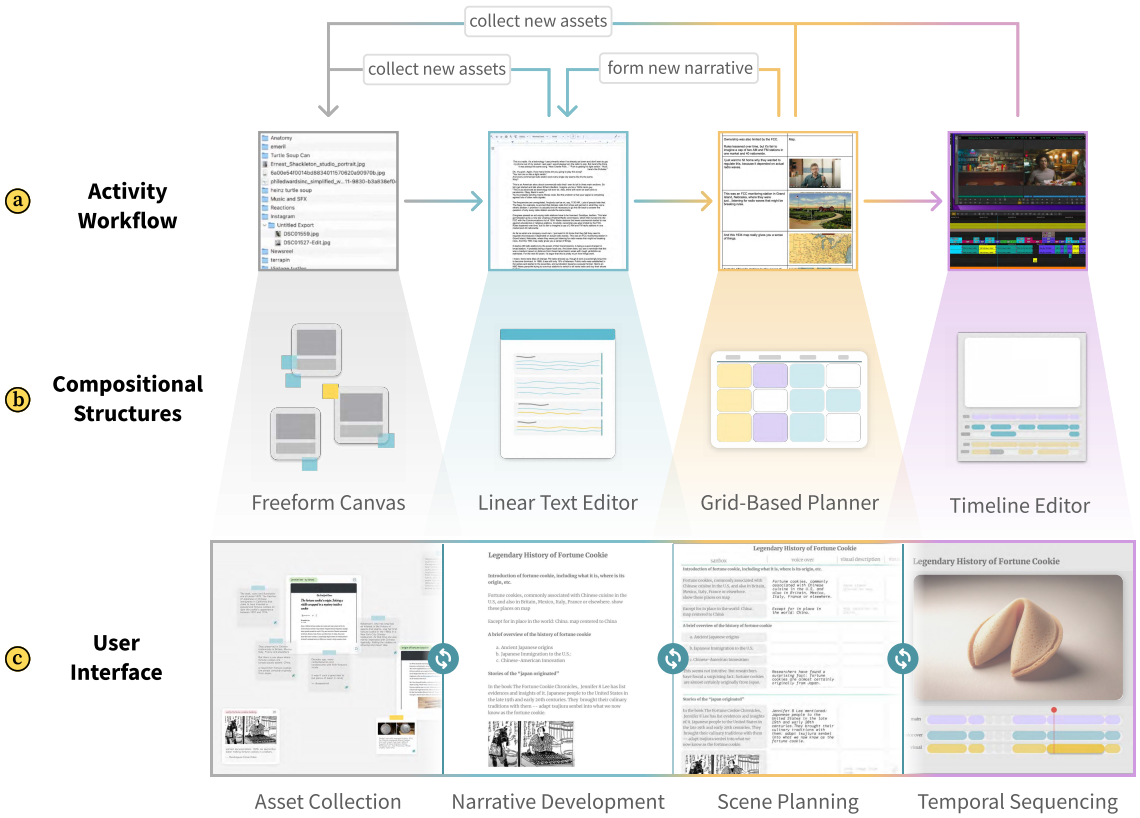


Figure 2: Mappings between the compositional structures identified in the video creation workflow and the VideOrigami's user interface. (a) Workflow revolving around the compositional structures; (b) the underlying compositional structures; (c) four views in VideOrigami's user interface maps to a corresponding composition structure

4.2.3 Scene Planning with Grid-Based Editor. "I need to put everything together, and see whether the visual goes well with the text, whether the temporal sequence makes sense."(E1) Creators develop scene structures by examining the narrative and congruence aspects simultaneously: they experiment and specify the arrangement of materials to ensure visual and temporal coherence both within and across the scenes. A variety of grid-based structures are utilized, such as two-column scripts (E2, E5) and storyboards (E1, E4). For example, two-column scripts allow creators to arrange the narrative sequence in rows and organize the materials within a scene in columns (e.g., voiceover and desired visuals). This stage often features a "random filling" pattern, as creators may have incomplete ideas or uncertainty in certain parts of a scene, such as knowing the visual but not the voiceover, or vice versa. The grid structure also provides a clear overview of the creation progress—unfilled cells serve as visual indicators of incomplete elements (E2, E5).

4.2.4 Spatial/Temporal Arrangement and Preview using Timeline-Based Editor. "You have been creating it as a creator, and now you are watching it as a viewer when putting them together." (E5) The timeline structure provides creators with fine-grained controls to refine the temporal sequence of the video, which can range from small pacing adjustments like trimming a clip to changing the clip sequences (E1, E2, E5). All timeline-based editors also provide controls for adjusting the placements of visuals for a selected time frame. At this stage, creators constantly assess the effectiveness of the storytelling by previewing the assembled parts. This iterative process of a perceptual reasoning is critical, as E3 explains: *"Only when you see it do you realize it's not what you imagined."*.

4.3 Desired Interconnections Across Structures

The current digital environment discourages iteration across structures, as the lack of synchronization between them often results in a high cost of context switching. This often incentivizes creators to confine their iterations to the single structure they are currently working in, rather than leveraging the one that would be most effective for the task. For example, all creators we interviewed noted that once they transition to working on the timeline, they "never go back"(E1), even when certain structural changes could benefit from using the narrative editor or scene planner (E2). Below, we summarize a set of cross-structure interconnections desired in the video creation workflows.

4.3.1 Collect and Refer to Materials Anytime. Creators often need to cross-reference assets while working on different structures at any workflow stage. E2 highlighted the frustration of repeatedly searching through scattered file folders to locate assets, which are often organized differently depending on tasks such as writing

or timeline editing. They also mentioned that multiple passes are needed to ensure no assets are overlooked, consuming up to half a day. Systems should provide a centralized collection of assets, allowing creators to freely add assets within any structure and easily reference them when working on different structures.

4.3.2 Develop Cohesive Narrative from Fragmented Notes. During asset collection, creators often generate fragmented ideas or notes that do not immediately fit into the narrative. At times, they may struggle to incorporate some interesting excerpts into the existing storyline (E3, E4). This suggests that systems should facilitate the quick integration of fragmented ideas into the narrative while ensuring narrative coherence.

4.3.3 Provide a Warm Start for Developing the Scene Structure. To develop the scene structure, creators often need to manually locate and transfer many materials into the appropriate categories (e.g., specific rows or columns), such as copy-pasting paragraphs from the narrative to different cells and inserting images and clips. This "cold start" process can be tedious and discourage them from using the scene structures (E1). Systems should automate the initial organization of content into scene structures, providing a "warm start" that reduces manual effort and streamlines the workflow.

4.3.4 Enable Granular and Structural Adjustments for Temporal Sequencing. When adjusting pacing, creators often have diverse needs for corresponding edits: they require fine-grained controls, such as precise timing adjustments, which should be reflected across other structures to ensure these structures remain reusable for relevant tasks; they may also make broader structural changes to the sequence, which should be supported in alternative structures and automatically synchronized with the timeline. Systems should accommodate both granular and structural adjustments, ensuring synchronization across all relevant compositional structures.

5 VideOrigami: a Human-AI Video Co-Creation Environment

We develop VideOrigami by following the proposed design process, including surfacing the compositional structures in one unified space [ST2], defining their synchronization [ST3], and infusing AI within and across structures [ST4]. Section 5.1 explains the individual structures and AI-driven generation to support their completion; Section 5.2 covers synchronization across the structures and AI implementation; and Section 5.3 presents a scenario illustrating the creation of a video within the co-creation workspace VideOrigami.

5.1 Compositional Structures and Within-Structure Generations

We formally define the four compositional structures we identified: Freeform Canvas, Narrative Editor, Grid-based Scene Planner, and Timeline Editor (Fig. 2c)—along with their individual components, organization rules, and functional affordances.

The Freeform Canvas leverages spatial organization to facilitate asset collection and exploration. VideOrigami's implementation supports three types of nodes: *Asset Nodes* (Fig. 3a) for uploading media, importing web content, or generating visual content (Fig. 3b). Each asset node has associated *Note Nodes*, enabling creators to record information related to the assets when making sense of them. Users can manually edit notes or generate them using queries to extract specific content from the assets. The *Prompt Nodes* (Fig. 3c) are used for generating certain parts of the video (as further elaborated in Section 5.3).

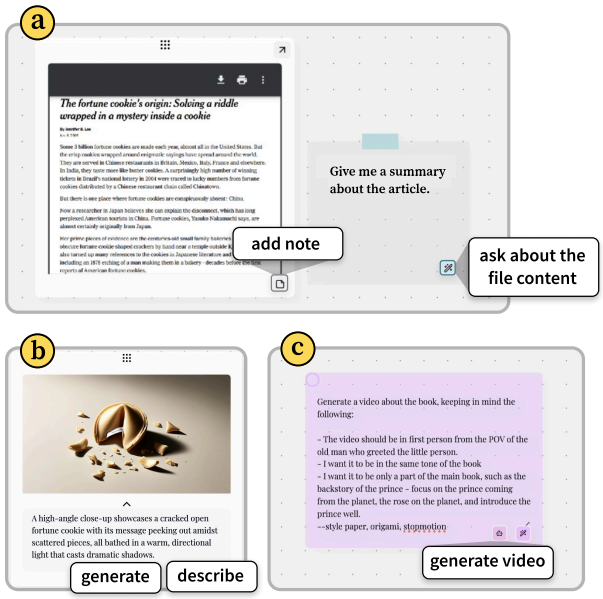


Figure 3: Nodes supported in the freeform canvas structure.

The Narrative Editor that we utilize is a linear block-based text editor that consists of two block types: section blocks, which define high-level section headings to guide the overarching structure of the video, and paragraph blocks, which contain individual talking points within each section. Users can manually edit each block and use LLMs to generate section headings to outline the narrative or talking points within specific sections. The generation process considers narrative cohesion and the relevance of each talking point within its context.

The Grid-Based Scene Planner employs columns and rows to plan different elements in a scene chronologically. The grid-based scene planner also defines four types of columns: the *Storyline* column houses the talking points; the *Script* column contains the transcript for each scene; the *Visual Description* column describes what visuals to show in each scene; and the *Visual Preview* column shows the visual assets to be included in the video. Besides basic operations such as adding, deleting, or shuffling rows and columns, users can populate each cell of the grid manually or generate content based on the context provided by existing rows and columns.

The Timeline Editor allows creators to preview the video and adjust its pacing. It organizes content into three types of tracks: audio, visual, and caption tracks. Each track comprises sequences of snippets as the smallest manipulable units in the timeline. While we did not implement AI features within the timeline, some techniques explored in prior work, such as aligning visual beats with audio [27] and transition suggestions [88] can be incorporated.

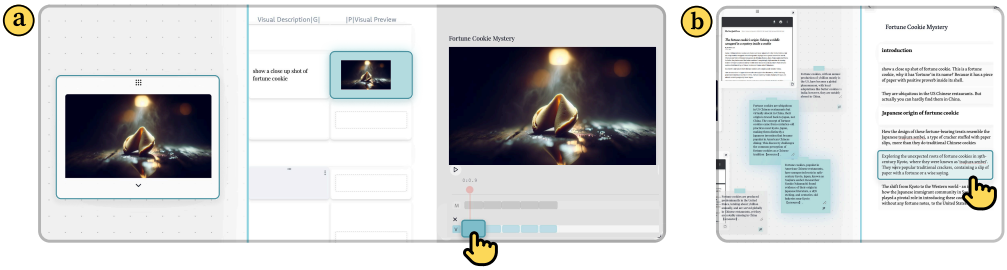


Figure 4: Synchronized highlighting between different structures. (a) When the user clicks on a segment in the Timeline Editor, the corresponding Grid cell is highlighted, and the node is centered on Canvas. (b) hovers hover a talking point in the Narrative Editor, VideOrigami highlights relevant note nodes in the Canvas.

5.2 Cross-Structure Synchronization and Transformation with AI

Informed by the desired interconnections (Section 4.3), we herein describe the cross-structure synchronization in VideOrigami in terms of the corresponding units, synchronization techniques, updating mechanisms, and the AI integration to support the synchronization.

5.2.1 Canvas ⇔ Other Structures (Assest Managment). The canvas serves as a centralized asset hub, where any item added to other structures is automatically added as an asset node. To facilitate referencing relevant assets, we implemented *bi-directional, automatic, synchronized highlighting*. Each canvas node corresponds to the smallest unit in other structures (i.e., a paragraph block in the narrative editor, a cell in the grid, or a snippet in the timeline). When a user edits a unit in another structure, related canvas nodes are dynamically highlighted (Fig. 4). In this process, AI is incorporated to actively calculate relevance using embedding vectors of the content. Conversely, when a user hovers over a Canvas node, related units in other structures are highlighted. This ensures that all important assets are effectively incorporated into the video.

5.2.2 Canvas ⇔ Narrative Editor (Narrative Development). To help users form cohesive narrative points from fragmented notes, we implemented *bi-directional, user-initiated, synchronized editing* between nodes in the canvas and the paragraph blocks in the narrative editor. When a user drags a note node into the narrative editor, VideOrigami leverages generative AI to transform the note content based on the drop target: dropping into a new block transforms its content into a new talking point with the existing narrative sequence as context; dropping into an existing block revises the content to integrate the note while preserving the original meaning (Fig. 5b). Conversely, when users drag a talking point into an empty canvas note node, VideOrigami extracts content relevant to that point from the note's associated asset content to further support this process (Fig. 5a).

5.2.3 Narrative Editor ⇒ Scene Planner (Scene Development). To provide an effective starting point for scene planning, we implemented *uni-directional, user-controlled, synchronized editing* from the narrative editor to the grid-based planner. The rows in the grid correspond to the different types of blocks in the narrative editor: a *section row* corresponds to a section block in the text editor (Fig. 6a); a *talking point row* corresponds to a paragraph block. When users modify content in the narrative editor, VideOrigami utilizes LLMs to categorize the content based on the grid's columns (e.g., visual descriptions or voiceover scripts) and maps it to the corresponding grid cells as suggested content (Fig. 6b). Users can press "Tab" to quickly accept the suggestion or overwrite them.

5.2.4 Scene Planner ⇔ Timeline Editor (Temporal Adjustment). To support both structural and fine-grain temporal adjustments, we implemented *bi-directional, synchronized editing* between the grid-based scene planner and timeline editor. The timeline can be seen as a transposed view of the scene planner, with each column corresponding to a track and individual grid cells representing specific time segments within those tracks. Users can drag columns to the timeline to form default tracks (e.g., the script column becomes an audio track, and the visual preview column becomes a visual track) (Fig. 7a). AI facilitates the transformation between grid cells and their corresponding time segments, with different updating mechanisms for each direction. Grid-to-timeline updates are *automatic*.

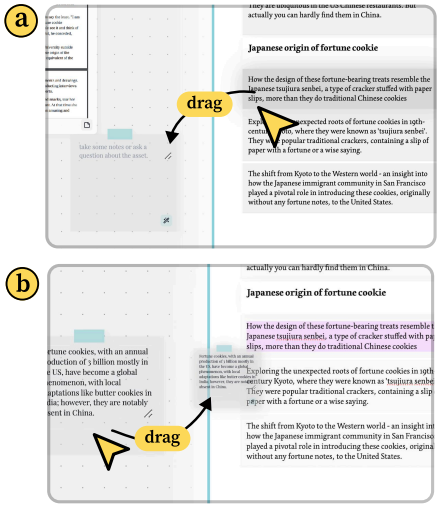


Figure 5: Transformation of information between Canvas and text editor. The user can (a) drag a talking point into an empty note to extract relevant content from the linked article; or (b) drag a note into the talking point to revise or create a new talking point based on extracted information from the article.

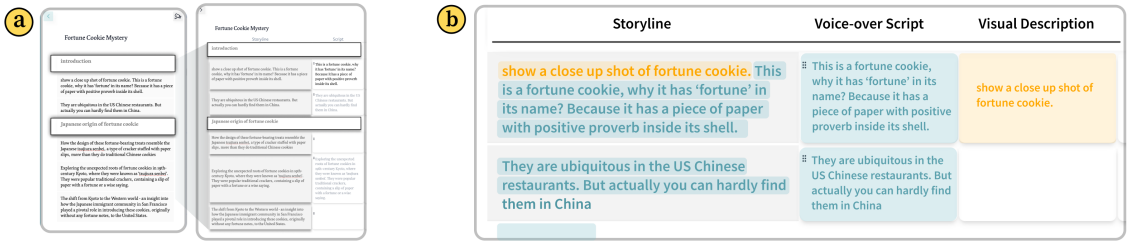


Figure 6: Transformation of information between the Narrative Editor and Grid. (a) Content in the text editor transforms into a storyline column in the Grid. (b) Population of the narrative content into Grid columns based on different column types.

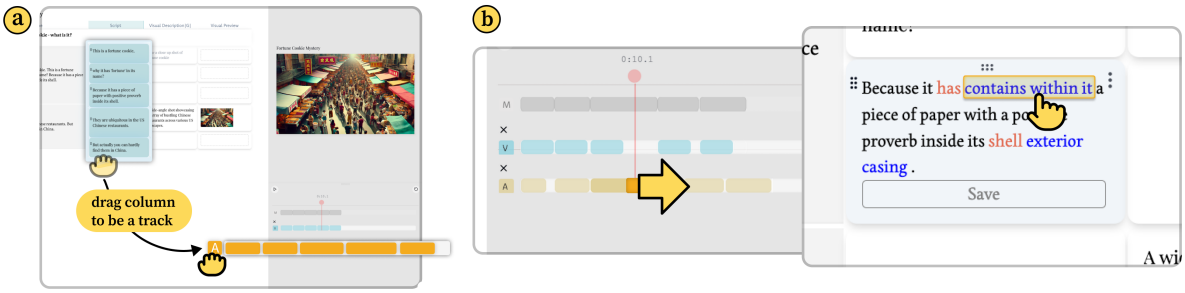


Figure 7: Transformation of information between Scene Planner and Timeline Editor. (a) Drag the script column to transform it into an audio track. (b) Fine-tune the time segments and adjust the corresponding script.

For example, extending a script or replacing a video clip will automatically adjust the corresponding time segments. Timeline-to-grid updates are *user-controlled*. For example, when a user shortens a time segment, VideOrigami suggests edits to the corresponding grid cell to align with the new timing while preserving the original meaning. Users can accept these suggestions or make their own edits (Fig. 7b).

5.3 System Walk-through

We walk through a scenario where a creator, Millie, uses VideOrigami to make an explainer video about "The Mystery of Fortune Cookies". Millie first imports some assets she has collected about fortune cookies into VideOrigami's canvas (Fig. 8a). Using the embedded AI functionality in the canvas, she explores the materials and conceptualizes two high-level sections for the video: *"The Origins"* and *"Fun Facts"* and adds them to the narrative editor.

5.3.1 Bottom-Up Creation. Millie wants to open the video with a familiar image, so she writes "*a close-up shot of a fortune cookie*" as the opening sentence. When transitioning to its Japanese origins, she recalls an article but forgets some details. VideOrigami highlights relevant notes in the canvas as reference (Fig. 8b). To seamlessly incorporate the facts, Millie drags a note into the paragraph, transforming it into a talking point while ensuring a cohesive narrative flow that integrates the new information (Fig. 5b). Satisfied with the story, Millie moves to the scene planner by clicking on the toggle located on the top right corner of the editor (Fig. 8c). VideOrigami transforms the content in the narrative editor to automatically populate the scene planner (Fig. 6), giving Millie an effortless first draft. With the synchronization, she also gets a rough cut in the timeline editor with a generated voiceover based on the script. Millie adds visuals, previews the video, and refines the scenes by iterating between the scene planner and the timeline editor.

5.3.2 Top-Down Creation. For the second section, Millie starts by prompting VideOrigami to generate the entire section first (Fig. 8d). Instead of directly presenting her a video cut, VideOrigami follows the structures and populates talking points, voiceovers, visuals, and time segments within each structure progressively. This allows Millie to review and refine as the video takes shape. She can reorganize sequences in the narrative editor to explore different narrative flows (Fig. 8e). As changes propagate across different structures, she can easily make further edits or preview the new version. Finally, Millie drags her own voiceover recording into the timeline to replace the generated audio (Fig. 8f). VideOrigami stores it as an asset in the canvas, as well as intelligently aligns her audio with the existing script in the grid, and adjusts the timing of each scene (Fig. 8g).

With the help of VideOrigami, Millie successfully creates her first cut of the explainer video. VideOrigami yields not only the output video, but also the artifacts of the process: the assets and notes on the canvas, the narrative in the editor, and the scene plan in the grid. These artifacts assist Millie in verifying sources and making future edits.

5.4 Implementation Details

VideOrigami is built using TypeScript with React for the front end, Zustand for state management across different views, MongoDB for the database, and a Python back-end. AI features include OpenAI's GPT-4 API for text generation, DALL·E 3 for image generation, and Whisper 1 for converting text to voiceover. A detailed description of the implementation is included in Appendix A.4.

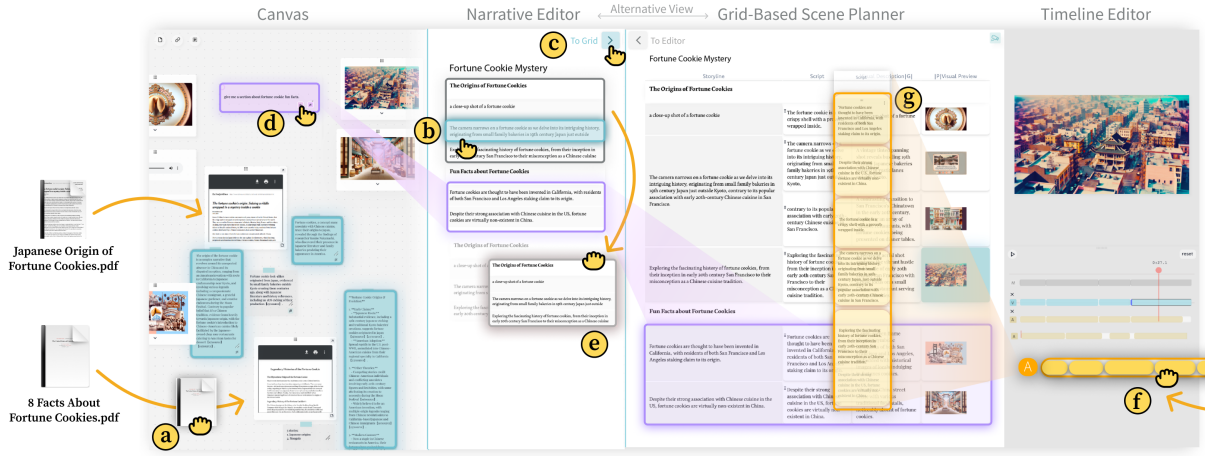


Figure 8: A user planning and creating a video using VideOrigami. The user can import external assets to the canvas (a); while the user is writing in the Narrative Editor, relevant notes are highlighted (b); the Narrative Editor expands and transforms into the Grid-Based Scene planner (c); the Prompt Node enables the user to generate a section of the narrative automatically (d); the user can re-arrange sections of the narrative by dragging and dropping in the Narrative Editor (e); the Timeline Editor enables users to import audio files (f) and automatically align the imported audio with the existing script (g).

6 Evaluation

We evaluated VideOrigami with a user study to investigate whether the design approach has resulted in an effective co-creation environment by answering the following research questions:

RQ1 Whether the aggregated compositional structures can serve as an effective common ground for human-AI collaboration;
RQ2 Whether AI can reduce challenges associated with completing and synchronizing the compositional structures;
RQ3 Whether there are new workflows and usage patterns enabled from this design approach;
RQ4 What tensions may arise in the co-creation environment?

6.1 Study Procedure

We evaluated VideOrigami with different creator profiles (i.e., experts and novices) and different creation workflows (i.e., human creates and AI synchronizes as well as AI creates and human refines) to more broadly evaluate the environment's effectiveness. The inherent complexities of video creation, coupled with the learning curve associated with a new authoring interface and its various features, resulted in the entire study exceeding 3 hours. Therefore, we broke the study into two parts. In the first part, all participants were asked to create a video with a bottom-up approach (i.e., human creates and AI synchronizes). In the second part, participants started with a single prompt to create a video, after which they evaluated and iterated upon it to achieve their desired outcome.

We recruited 6 novices (N1-N6, 3 female, 3 male), and 4 experts (E1-E4, 3 female, 1 male) for the study. Novices self-reported having limited knowledge of video creation tools and techniques and have created less than ten videos. All experts had at least two years of video editing experience and published videos either monthly (E4) or once every few months (E1-3). All participants attended Part 1 of the study. Novices and experts were compensated with $40 and $100, respectively, for their participation in Part 1 (2 hours). Three novices (N1-3) and three experts (E1-3) attended Part 2 (1.5 hours), and received $25 and $50, respectively.

6.1.1 Part 1: Bottom-Up Creation. In this 120-minute study, participants (6 novices, 4 experts) were tasked to create a 30-second video about fortune cookie origins from scratch using our system. Before the study, participants were given two articles to familiarize themselves with the video topic. Seven studies were conducted in-person and three remotely via Zoom. All participants accessed VideOrigami through a web browser.

Introduction and System Walk-through (~40 minutes). The experimenter first introduced participants to VideOrigami and interviewed them about their experience with video creation and generative AI. Next, the experimenter walked the participants through VideOrigami's features by guiding them in creating the first half of the video (introduction to fortune cookies). During the walk-through, the experimenter explained each interaction and asked participants to perform specific actions.

Creation task (~40 minutes). Participants were asked to complete the second half of the video (focusing on the cookie's Japanese origins). The system was pre-loaded with assets they could optionally use during their creation process.

Iteration task (~15 minutes). Participants were given an additional article describing stories of fortune cookies. They were instructed to extract narrative points from the article and revise the video to add at least one narrative point.

Questionnaire and Post Interview (~25 minutes). After completing all the tasks, participants filled out a questionnaire about the usefulness of VideOrigami's concept, features, and their experience, followed by a semi-structured interview to gather further insights.

6.1.2 Part 2: Top-Down Creation. In this study, participants created another 30-second video. To allow the participants to easily evaluate AI-generated content (e.g., narrative, images, video sequences), we asked the participants to prepare materials they were familiar with,

such as a novel, article, or blog post, as the information source. Three studies were conducted in-person and 3 via Zoom.

Task Introduction and System Revisit (~10 minutes). Participants were first given a general introduction to the task, followed by a quick walk-through of the system to serve as a refresher.

Practice Task (~10 minutes). Participants were instructed to use AI to generate a video with the content they provided by writing a simple prompt such as 'create a video'. The goal of this step was to help them get a sense of the generation process and the generated video.

Creation Task (~45 minutes). Participants were asked to write more detailed prompts to generate a video and then refine the generated video through iterative adjustments using the system.

Questionnaire and Post-study Interview (~20 minutes). After completing all the tasks, participants completed a questionnaire about the usefulness of structures in iteration and comprehension, followed by a semi-structured interview to gather further insights and compare their experience with the bottom-up approach.

7 Findings

Results from the questionnaires and interviews provide evidence of using compositional structures to ground human-AI collaboration and facilitate fluid iteration of various aspects of the video. The compositional structures and generative AI together yield a new cost structure [70] for video creation, enabling new workflows and unveiling exciting research questions that require further investigation. We have included a few samples of the outputs created during our user study in the Appendix A.1, and more results created during our user study can be found in this gallery.[1]

In the sections that follow, we describe our findings in terms of the four research questions and the general implications we draw from these findings that can potentially apply to other creation domains in the use of compositional structures for human-AI co-creation environments.

7.1 Compositional Structures as Strong Foundation for Collaboration (RQ1)

7.1.1 Effective Representation of the Whole Picture. By aggregating multiple compositional structures inherent in video creation together, VideOrigami preserves and visualizes the entire creation context and progress. Experts found the structures familiar, and novices found the system instructive and could help them develop an understanding of video creation and "easily get hands-on" (N3) using each structure to compose different aspects of a video.

> The way you are putting those things here has a structured layout which helps me to understand how the video is composed. (N3)

Additionally, both experts and novices reported that the system enabled them to stay aware and oriented in a typically messy process. They could "understand what's going on" (E3) and "pick up wherever they want" (N5). Interestingly, while we were targeting human-AI collaboration, multiple participants commented on the system's suitability for human-human collaborations, indicating VideOrigami's effectiveness as a general collaboration platform.

> I like that the context is always in front of me so I'm not blinded by anything. (N2)

7.1.2 Facilitating Understanding Generated Content and Generation Capabilities. Participants found these structures particularly effective in helping them comprehend AI-generated content and identify issues in AI generation. E2 mentioned that the structures allowed them to "directly make sense of the generated results". E1 mentioned that by reading the generated section headings in the narrative editor, they could *"quickly know how AI suggests the story should go"*. N2 noted on the grid structure in the scene editor *"gives me more transparency about what goes into each frame"*.

> I like the fact that I know what the breakdown is. I'm seeing what the script is, and I'm already thinking about what the visuals seem to look like. And then I can quickly spot something that does not match my expectations and know where I should change. (N2)

Beyond helping creators understand what AI is generating, structures also facilitate the understanding of what AI can generate, offering creators a glimpse of achievable results.

> I see the Rose and the Bob in the image, but that has nothing to do with the text in this area. So it tells me that there's some global understanding of things here... that makes me feel confident about realizing there is some cohesion to this, and I can maybe, through better prompting, get somewhere. (N1)

The ease of staying aware of the whole picture of video creation and comprehending AI's work and capabilities as a collaborator bolstered participants' awareness and assurance in collaborating with AI on the video creation task.

7.2 Leveraging Generative AI to Complete and Synchronize Compositional Structures (RQ2)

All participants reported that the synchronization (Mean=5, SD=0) and generative function (expert M=5, SD=0; novice M=4.83, SD=0.41) increased their productivity by helping them quickly get to a rough cut from a blank canvas.

7.2.1 Shortening the Path to the First Rough Cut. As mentioned in the formative study with video creators, a unique challenge of video creation is the lengthy process required to reach the first rough-cut preview compared to other tasks such as writing or graphical design. Creators must constantly speculate about the final audio-visual outcome while collecting assets, developing narratives, and planning the scenes until the very last stage. Yet, they are frequently surprised by the significant gap between their expectations and the actual outcome. However, the time constraint at the last stage of the workflow leaves little room for major structural changes. A significant benefit of the unprecedented speed enabled by AI is the ability to quickly create a rough cut to examine whether the final result matches their expectation. This was found particularly useful by participants, as it allowed them to quickly get a sense of whether the narrative is effective.

[1] VideOrigami User Study Results: https://videorigami-userstudy.netlify.app/

> I think it was a good start because usually, I feel the most difficult part is to get a timeline out. With a timeline, it will be much easier to tweak. (E2)
>
> I'm struggling to even build that initial structure; that's when I would like to have this as my starting point and then build on it. (N2)

7.2.2 Parallel Development and Iteration of Multiple Compositional Structures. Different from what we found in the formative study, where creators tend to complete one structure as much as possible before they move to the next, during both parts of our study, we observed a frequent switch across the structures throughout the creation process. At the initial stage, N1 started with exploring the canvas, while N5 started by generating multiple talking points in the narrative editor and brainstormed from there. N2, and E3, on the other hand, switched frequently between the narrative editor and scene editor for narrative development. After they had parts of the story, some creators (N4, E4) previewed the videos frequently when making iterative edits, while other creators (N3, E1) frequently switched between the scene editor and canvas. This usage pattern proved creators' desire for and VideOrigami's ability to support inspecting and manipulating different compositional structures during creation.

All creators strongly agreed that the synchronizations made their creation processes more productive by reducing the effort of *"moving things around"* (E2). Some commented that this also made them more creative by allowing them to focus on the creative aspect of the process (N5, E1, E2).

7.2.3 AI that Went Unnoticed. Because the synchronization among the compositional structures follows the constraints of the structures, they are less error-prone than generating an entire narrative from talking points or generating the visuals based on text descriptions. As a result, the intelligent synchronization often went unnoticed. In the current system, synchronization is approached conservatively, where we use visualizations to show connections and update the content in a suggestive manner. However, feedback from participants indicates a desire for more proactive synchronization. For example, when switching between narrative editor and scene planner, multiple creators expressed preferences for automatically transforming their storyline into a revised version of the voice-overs or visual descriptions.

> The live synchronization, it just seemed so natural like that's how it should be, I did not realize I was using it all the time. (N1)
>
> Since I need to review it anyways, I would prefer to have it [AI] directly transform my wording. (N3)

7.3 Emergent Workflows Due to Shifts in the Cost Structures (RQ3)

In addition to the parallel development and iteration of multiple compositional structures, we observed other novel workflows due to the shifts in the cost of content creation and evaluation process.

7.3.1 Grid-Based Scene Structure as the Main Playground. Despite the different usage patterns across participants, the grid-based scene planner stood out as the pivotal structure for the entire creation process when using VideOrigami. E1 described it as their *"main battlefield"*. This is different from what we found in the formative study, where creators usually spend most of their time on the narrative and timeline, as they did not want to be constrained by a rigid grid structure and considered completing a detailed scene planning in the grid too much work.

With VideOrigami, the grid-based scene structure not only organizes and displays all the materials used in the video but also connects with the narrative editor and the timeline. Participants realized they could command all the materials, and any changes made in the scene planner could propagate to the narrative structure and timeline. As a result, they felt comfortable entering the grid much earlier in the creation workflow and sticking with it.

7.3.2 Hidden Cost of Generation. With significantly lowered costs of generating and synchronizing the structures, we expected participants to leverage these capabilities to explore different video ideas and perform more large-scale revisions. However, these happened less than what we expected. This could be because of the constraints inherent in the study settings, such as the limited time participants had to explore different ideas and their limited commitment to the outcome. Nevertheless, we also observed evidence of how the low cost of the interaction and automation techniques disguise other costs in the entire creative activity.

In the first part of the study, participants were requested to develop the video piece by piece. Because of the low cost of using AI to generate content, we observed participants using AI to quickly fill the narrative and scene structure without giving too much thought. In particular, participants reported that filling in the empty cells in the grid was very tempting, as they could quickly *"have something look good and get a story across"*(N1). As a result, while some of the generated images and text were not ideal, they often opted for filling the entire structure instead of spending effort refining specific elements. Yet, the cost of structural revision of the video quickly rises as the video becomes more complete, discouraging participants from large-scale iterations.

In the second part of the study, given the low cost of generating a video, we expected participants to explore different ideas by prompting several rounds. However, only 2 participants (N3, E3) made section-level changes to the generated storyline. We observed several reasons. First, participants often found the generated content good enough as a cohesive piece, albeit different from what they expected (E1, E3). More importantly, while the cost of generating a video was significantly lower, the cost of assessing a video remained high, including the time spent waiting for the video generation, reviewing the video, comprehending the narrative, and reviewing the visual prompts. As a result, participants were incentivized to accept a good enough generation and only perform small iterations.

> Now that I see it. I feel like alien invasion is a good starting point because it captures people's attention. Ok, I actually think the first part [of AI-generated narrative] makes sense to me now. (E1)
>
> I was mainly looking at how the storyline was generated and compared it to what I was expecting… It's interesting that it only did the beginning of my prompt, but I like the storyline it gave. It makes sense.

> So I'm not going to tweak anything with the storyline, mainly the script and visual description. (E3)

These findings suggest that while the combination of compositional structures and generative AI reduces the cost of manual operation, it does not magically reduce all costs but may also introduce new ones. More specifically, our findings indicate that lower operation-wise costs do not necessarily lead to increased iterations. The high-fidelity designs that AI generates create the impression of completeness, which may discourage further iterations.

7.4 Tensions of Creative Expressions (RQ4)

While all participants mentioned the generative functions for filling the structures made them feel productive, their opinions on whether this made them creative were mixed. Some participants appreciated that the generated results could provide them with ideas they had not thought of, hence *"enlarged creativity realm"* (E1). Others found the generated results impeded their creativity, despite knowing they could create everything manually. Novices and experts also differ in their perceptions. With a 5-point Likert scale, novices (M=4.5, SD=0.84) reported feeling more creative when using the generative functions compared to experts (M=3.5, SD=0.96); novices (M=4.33, SD=0.52) also felt a greater sense of transparency and control over the AI functionalities compared to experts (M=3.5, SD=1).

7.4.1 Overshoot, Undershoot, and Sweetspot of Generation Fidelity. Some participants found the high-fidelity content generated by AI overwhelming and impeded their creative thinking. *"Sometimes, I have a vague idea in mind, but the generated visual is so detailed that, upon seeing it, I find my thoughts blocked"* (E4). This indicates that the concrete nature of generated visuals can overshadow nascent ideas. On the other hand, creators felt frustrated and a sense of *"lost control"* when AI failed to generate their envisioned visuals. Interestingly, the visual description generated by AI, which described the intended visual for a scene, was positively received by all participants. We observed both novices and experts, regardless of whether the visuals matched their expectations, frequently inspect AI-generated descriptions of the visual content, as they found AI's description of suitable visuals, compared to actual visual content, both informative enough and leaving room for their own ideas.

7.4.2 Creative Expression with Prompts. We observed that participants' expertise in prompting significantly impacted their creation experience. Participants who lacked prompting experience often got confused about why AI produced certain results, whereas participants who are experienced with generative AI tools could better comprehend issues in the generated results.

> By comparing the visual descriptions and the images in these rows, I know it got confused by this abbreviation. (N3, comprehending AI's mistakes)

> Originally, I thought that the script would be the poem, line by line, right? But I don't know where this is coming from. It's like a third-person perspective. Just wonder what's the logic here.. (E1, confused by the AI-generated storyline)

Participants' abilities to comprehend AI-generated content and to express their desired outcomes in prompts directly affected the effectiveness of their iterations. Being skillful at prompting enhanced participants' sense of control during the collaboration, broadened their horizon of what content was achievable, and maintained their creative engagement with the creation process.

7.4.3 Contrast between Novices and Experts . The contrasts between novices' and experts' perceptions towards AI-generated content were also observed in our study. Experts, such as E2, reported concerns about the decline in their engagement with the creative process. Novices, on the other hand, were much more receptive to AI-generated content, as AI empowered them to create content beyond their abilities.

> It's so fast to generate results, and the result makes sense. I found myself stopping thinking during the process... If I am emotionally invested in a project, I might be hesitant to use AI, so I won't lose commitment to it. (E2)

> Because I know that I lack the ability to do a lot of things ... so this (AI) really gets me going. (N1)

As discussed in the previous section, participants' expertise in prompting AI significantly affected their creation outcomes and experiences. Within the research team, we found videos created by novices with significantly more prompting experiences exhibited higher quality than those created by expert participants during our study. It is imperative to note, however, that this observation does not serve as evidence suggesting that novices with proficiency in prompting are capable of creating content of a higher caliber than experts. Experts, by virtue of their extensive knowledge and understanding of the principles underlying the production of high-quality videos, maintain a distinct advantage. Nonetheless, the findings of our study suggest that expertise in prompt engineering can, to a certain degree, reduce the gap between novices and experts in creating creative content.

7.5 Summary

Findings from the user evaluation of VideOrigami provide evidence of the effectiveness of the design approach of employing compositional structures as the foundation of the human-AI co-creation environment. Specifically, we found the aggregated and interconnected compositional structures enabled creators to stay oriented throughout the creation process and facilitated their understanding and control of AI generation. Participants acknowledged that they felt more productive with AI helping them complete and synchronize the structures. We also observed new workflows and costs associated with human-AI video co-creation, which we discuss further below.

8 Discussion and Future Work

The user evaluation allowed us to understand the effectiveness of the human-AI video co-creation environment created with our design approach. We first discuss how the cost structure of the video creation activity is shifted and then discuss the implication of the design approach by situating it in the broad scope of information work and activity-centered information spaces.

8.1 New Costs in Human-AI Video Co-Creation and Beyond

Besides the commonly recognized challenge of prompting AI to generate the desired content, our study uncovered three new types of costs. While these costs were observed in the context of video creation, they are broadly applicable to other domains within human-AI co-creation.

8.1.1 Evaluation Cost Inherent in Content. The high cost of evaluating a generated video became prominent in our study, which made participants reluctant to explore different video ideas. Participants not only had to examine the final generated content but also the underlying narrative and the reasons for AI generating the visuals for specific scenes. Among the various types of AI-generated content, images only require a glance to determine their suitability, whereas other types of content, such as text and video, require significant cognitive effort to consume and evaluate. To reduce the evaluation cost of text, research has explored transferring text to diagrams or summarizing the text to facilitate consumption of a large amount of text [36, 48].

Future research should investigate how to apply these ideas to reduce the evaluation cost of video and other content emphasizing the congruent, narrative, and temporal aspects, as these aspects inherently demand significant cognitive loads to evaluate.

8.1.2 Harms of High-fidelity Content. AI-generated content is often of high fidelity, which can be harmful to early-stage design, as the high-fidelity content pushed the participants toward early convergence when they could benefit from more diverse ideas [29, 30, 76, 83, 85]. One approach to mitigate this problem is to present users with multiple options, which was found to encourage exploration in design [30]. However, as mentioned above, evaluating a generated video incurs significant cost, let alone reviewing and comparing multiple options. Another option is instructing AI to purposefully generate low-fidelity prototypes to avoid nudging users to converge too early or generating design space to encourage exploration [76].

Future research should explore combining these approaches while developing new strategies to mitigate the potential harm the high-fidelity content may pose to users' agency and creativity.

8.1.3 Fear of Losing Previous Versions. Popular AI tools such as, ChatGPT and Midjourney, utilize chat as the primary interaction mechanism. A benefit of the chat mechanism is that all previous prompts and results are automatically preserved, enabling users to recover to a previous version easily. VideOrigami avoided using chat due to natural language's inefficiency in supporting flexible reference and manipulation of inherently spatial structures [63], and instead employed the direct manipulation paradigm [47]. The downside, however, is that every edit operation is destructive. While undo/redo and version control can be provided, they typically do not make history as directly visible and easily retrievable as the chat mechanism. We found this contributed to participants' reluctance to iterative prompt engineering, which is often seen in other AI-assisted workflows. Future work will explore how to incorporate the version control mechanism to meet the same level of visibility and accessibility.

8.2 Generalizability of the Design Approach

Synthesized from the literature across diverse domains, our design approach—identifying, designing, synchronizing, and integrating compositional structures—provides a robust framework for developing human-AI co-creation environments and is inherently generalizable to the domains we surveyed, including writing, music, podcast, and interactive media. We believe the design approach can also generalize to other types of content of a compositional nature, such as game and VR scene development. We discuss key design considerations when applying the design approach.

8.2.1 Devising New Compositional Structures. Our design approach emphasizes identifying compositional structures within existing workflows. However, we recognize that novel compositional structures can often bring significant benefits to the content creation workflow [31, 67, 84, 106]. For example, time-aligned transcript [99] has significantly reduced the editing effort for audio and video content [68, 84, 101] by enabling creators to more easily attend to the temporal and congruent aspects of the content. When designing co-creation spaces, it may be fruitful to consider what content aspects receive inadequate support and devise new compositional structures to fill the gap.

8.2.2 Supporting Freeform Exploration and End-user Customizable Structures. Freeform canvas is a common structure used in all domains we surveyed, despite that it does not enforce specific compositional rules. This is because freeform canvas enables creators to experiment freely with different compositions, especially during exploration. Structures that can be freely created and customized based on end-users' needs can also be beneficial. This may require the development of lower-level primitives with meta-compositional rules that describe how compositional structures themselves should be composed.

8.2.3 Managing Multiple Compositional Structures within Unified Workspaces. An interface design challenge is how to effectively organize and integrate multiple compositional structures within the interface without incurring significant navigation and interaction costs. VideOrigami, for example, supports the merging of narrative editor and scene planner to reduce screen clutter and management costs. Domain-specific solutions may need to be devised when developing the co-creation space for other domains.

8.2.4 Understanding the Shifts of Cost with Compositional Structures. Our study revealed that creators' reliance on the structures may shift as they are integrated. For example, in our formative study, some creators skipped using the grid-based structures due to the high cost of developing and maintaining the structures. However, participants in our user study found the grid-based structure pivotal as it allowed them to engage with multiple content aspects. Therefore, designers of co-creation spaces need to understand the shifts of cost associated with integrating and synchronizing the compositional structures and their implications on users' workflows, and design the interface accordingly.

8.3 Beyond Content Creation and Compositional Structures

A clear next step of this work is to apply the design approach to other domains to further verify its generalizability. Beyond that, compositional structures are not the only structures employed in content creation. Structures that enable exploration of the design space, comparison of options, and tracking of changes are equally important in content creation workflows. Additionally, tasks like planning and decision-making often rely on various structures—for example, using tables to compare options during travel planning or data visualizations to facilitate filtering and ranking. A promising direction for future research is to develop a comprehensive taxonomy of these structures, along with methods for interconnecting them to allow seamless transformation of information across different structures.

The human-AI co-creation environments are *de facto* activity-centered information spaces that encompass various information structures and functionality traditionally distributed across individual applications [35]. Developing activity-centered, instead of application-centered, spaces has been a long-lasting endeavor in HCI [13, 89, 102]. While previous attempts at this vision have failed for various reasons, we believe pursuing human-AI co-creation environments is another promising attempt. Based on the taxonomy mentioned above, we will develop a structure library that enables developers to easily create such environments or enable AI to intelligently compose such environments based on users' tasks.

9 Conclusion

AI is transforming not only how information is generated but also the fundamental structure of the information environment. In this work, we present an initial exploration of a design approach for developing human-AI collaborative environments. Specifically, we propose integrating compositional structures of information activities and embedding AI within and across these structures to create a cohesive, intelligent collaborative environment. Our findings from a video co-creation environment developed using this approach demonstrate that such an environment helps creators remain oriented within the creation workflow, gain greater control and interpretability of AI generation, and flexibly interweave human-driven and AI-driven processes. Grounded in the compositional nature of complex information content, with video creation as a representative activity, we believe this approach has significant potential for broad application across various domains.

References

[1] Ableton. [n. d.]. Ableton - Music production with Live and Push. https://www.ableton.com/en/. Accessed: 2024-09-07.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[4] Adobe Inc. 2024. Adobe Audition. https://www.adobe.com/products/audition.html. Accessed: 2024-09-07.

[5] Adobe Inc. 2024. Adobe InDesign. https://www.adobe.com/products/indesign.html. Accessed: 2024-08-24.

[6] Adobe Inc. 2024. Adobe Premiere Pro. https://www.adobe.com/products/premiere.html. Accessed: 2024-08-24.

[7] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. 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, 1–13. doi:10.1145/3290605.3300233

[8] Apple Inc. 2024. *Logic Pro.* Accessed: 2024-09-07.

[9] Audacity Team. 2024. Audacity - Free, open source, cross-platform audio software. https://www.audacityteam.org/. Accessed: 2024-09-07.

[10] Benjamin Bach, Zezhong Wang, Matteo Farinella, Dave Murray-Rust, and Nathalie Henry Riche. 2018. Design Patterns for Data Comics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3173612

[11] Michel Beaudouin-Lafon. 2017. Towards Unified Principles of Interaction. In *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter* (Cagliari, Italy) *(CHItaly '17)*. Association for Computing Machinery, New York, NY, USA, Article 1, 2 pages. doi:10.1145/3125571.3125602

[12] Floraine Berthouzoz, Wilmot Li, and Maneesh Agrawala. 2012. Tools for placing cuts and transitions in interview video. *ACM Trans. Graph.* 31, 4, Article 67 (jul 2012), 8 pages. doi:10.1145/2185520.2185563

[13] Susanne Bodker. 2021. *Through the interface: A human activity approach to user interface design.* CRC Press.

[14] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[15] Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-image generation through interactive prompt exploration with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[16] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. 2024. Video generation models as world simulators. (2024). https://openai.com/research/video-generation-models-as-world-simulators

[17] Daniel Buschek. 2024. Collage is the New Writing: Exploring the Fragmentation of Text and User Interfaces in AI Tools. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 2719–2737. doi:10.1145/3643834.3660681

[18] Yining Cao, Jane L E, Zhutian Chen, and Haijun Xia. 2023. Dataparticles: Block-based and language-oriented authoring of animated unit visualizations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[19] Yining Cao, Rubaiat Habib Kazi, Li-Yi Wei, Deepali Aneja, and Haijun Xia. 2024. Elastica: Adaptive Live Augmented Presentations with Elastic Mappings Across Modalities. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.

[20] Vincent Cavez, Catherine Letondal, Emmanuel Pietriga, and Caroline Appert. 2024. Challenges of Music Score Writing and the Potentials of Interactive Surfaces *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 728, 16 pages. doi:10.1145/3613904.3642079

[21] Zhutian Chen and Haijun Xia. 2022. CrossData: Leveraging Text-Data Connections for Authoring Data Documents. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 95, 15 pages. doi:10.1145/3491102.3517485

[22] Peggy Chi, Tao Dong, Christian Frueh, Brian Colonna, Vivek Kwatra, and Irfan Essa. 2022. Synthesis-Assisted Video Prototyping From a Document. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–10.

[23] Peggy Chi, Zheng Sun, Katrina Panovich, and Irfan Essa. 2020. Automatic video creation from a web page. In *Proceedings of the 33rd annual ACM symposium on user interface software and technology*. 279–292.

[24] 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 *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 209, 19 pages. doi:10.1145/3491102.3501819

[25] Matthew Conlen, Megan Vo, Alan Tan, and Jeffrey Heer. 2021. Idyll Studio: A Structured Editor for Authoring Interactive & Data-Driven Articles. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3472749.3474731

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

Software and Technology (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 98, 13 pages. doi:10.1145/3526113.3545672

[27] Abe Davis and Maneesh Agrawala. 2018. Visual rhythm and beat. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–11.

[28] Descript. 2024. Descript: Audio and video editing software. https://www.descript.com Accessed: YYYY-MM-DD.

[29] Steven Dow, Julie Fortuna, Dan Schwartz, Beth Altringer, Daniel Schwartz, and Scott Klemmer. 2011. Prototyping dynamics: sharing multiple designs improves exploration, group rapport, and results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2807–2816. doi:10.1145/1978942.1979359

[30] Steven P. Dow, Alana Glassco, Jonathan Kass, Melissa Schwarz, Daniel L. Schwartz, and Scott R. Klemmer. 2011. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Trans. Comput.-Hum. Interact.* 17, 4, Article 18 (dec 2011), 24 pages. doi:10.1145/1879831.1879836

[31] Marianela Ciolfi Felice, Nolwenn Maudet, Wendy E. Mackay, and Michel Beaudouin-Lafon. 2016. Beyond Snapping: Persistent, Tweakable Alignment and Distribution with StickyLines. In *Symposium on User Interface Software and Technology (UIST)*.

[32] Syd Field. 2005. *Screenplay: The foundations of screenwriting*. Delta.

[33] Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College composition and communication* 32, 4 (1981), 365–387.

[34] Amy Rae Fox, Philip Guo, Clemens Nylandsted Klokmose, Peter Dalsgaard, Arvind Satyanarayan, Haijun Xia, and James D. Hollan. 2020. Towards a dynamic multiscale personal information space: beyond application and document centered views of information. In *Companion Proceedings of the 4th International Conference on Art, Science, and Engineering of Programming* (Porto, Portugal) *(Programming '20)*. Association for Computing Machinery, New York, NY, USA, 136–143. doi:10.1145/3397537.3397542

[35] Amy Rae Fox, Philip Guo, Clemens Nylandsted Klokmose, Peter Dalsgaard, Arvind Satyanarayan, Haijun Xia, and James D. Hollan. 2020. Towards a dynamic multiscale personal information space: beyond application and document centered views of information. In *Companion Proceedings of the 4th International Conference on Art, Science, and Engineering of Programming* (Porto, Portugal) *(Programming '20)*. Association for Computing Machinery, New York, NY, USA, 136–143. doi:10.1145/3397537.3397542

[36] Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review* 47, 1 (2017), 1–66.

[37] Jérémie Garcia, Louis Bigo, Antoine Spicher, and Wendy E. Mackay. 2013. PaperTonnetz: supporting music composition with interactive paper. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris, France) *(CHI EA '13)*. Association for Computing Machinery, New York, NY, USA, 3051–3054. doi:10.1145/2468356.2479608

[38] Jérémie Garcia, Theophanis Tsandilas, Carlos Agon, and Wendy Mackay. 2012. Interactive paper substrates to support musical creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1825–1828. doi:10.1145/2207676.2208316

[39] Jérémie Garcia, Theophanis Tsandilas, Carlos Agon, and Wendy E. Mackay. 2014. Structured observation with polyphony: a multifaceted tool for studying music composition. In *Proceedings of the 2014 Conference on Designing Interactive Systems* (Vancouver, BC, Canada) *(DIS '14)*. Association for Computing Machinery, New York, NY, USA, 199–208. doi:10.1145/2598510.2598512

[40] Han L. Han, Junhang Yu, Raphael Bournet, Alexandre Ciorascu, Wendy E. Mackay, and Michel Beaudouin-Lafon. 2022. Passages: Interacting with Text Across Documents. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 338, 17 pages. doi:10.1145/3491102.3502052

[41] John Hart. 2013. *The Art of the Storyboard: A filmmaker's introduction*. Routledge.

[42] Jeffrey Heer, Matthew Conlen, Vishal Devireddy, Tu Nguyen, and Joshua Horowitz. 2023. Living Papers: A Language Toolkit for Augmented Scholarly Communication. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 42, 13 pages. doi:10.1145/3586183.3606791

[43] Rorik Henrikson, Bruno Araujo, Fanny Chevalier, Karan Singh, and Ravin Balakrishnan. 2016. Multi-Device Storyboards for Cinematic Narratives in VR. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 787–796. doi:10.1145/2984511.2984539

[44] Rorik Henrikson, Bruno De Araujo, Fanny Chevalier, Karan Singh, and Ravin Balakrishnan. 2016. Storeoboard: Sketching Stereoscopic Storyboards. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4587–4598. doi:10.1145/2858036.2858079

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

[46] Bernd Huber, Hijung Valentina Shin, Bryan Russell, Oliver Wang, and Gautham J Mysore. 2019. B-script: Transcript-based b-roll video editing with recommendations. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[47] Edwin L Hutchins, James D Hollan, and Donald A Norman. 1985. Direct manipulation interfaces. *Human–computer interaction* 1, 4 (1985), 311–338.

[48] Peiling Jiang, Jude Rayan, Steven P. Dow, and Haijun Xia. 2023. Graphologue: Exploring Large Language Model Responses with Interactive Diagrams. In *Symposium on User Interface Software and Technology (UIST)*. https://doi.org/10.1145/3586183.3606737

[49] DaYe Kang, Tony Ho, Nicolai Marquardt, Bilge Mutlu, and Andrea Bianchi. 2021. ToonNote: Improving Communication in Computational Notebooks Using Interactive Data Comics. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 727, 14 pages. doi:10.1145/3411764.3445434

[50] 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.

[51] Nam Wook Kim, Nathalie Henry Riche, Benjamin Bach, Guanpeng Xu, Matthew Brehmer, Ken Hinckley, Michel Pahud, Haijun Xia, Michael J. McGuffin, and Hanspeter Pfister. 2019. DataToon: Drawing Dynamic Network Comics With Pen + Touch Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300335

[52] Cynthia L King. 2012. Reverse outlining: A method for effective revision of document structure. *IEEE Transactions on Professional Communication* 55, 3 (2012), 254–261.

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

[54] Clemens N Klokmose, James R Eagan, Siemen Baader, Wendy Mackay, and Michel Beaudouin-Lafon. 2015. Webstrates: shareable dynamic media. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. 280–290.

[55] Wilmot Li, Maneesh Agrawala, Brian Curless, and David Salesin. 2008. Automated generation of interactive 3D exploded view diagrams. *ACM Trans. Graph.* 27, 3 (aug 2008), 1–7. doi:10.1145/1360612.1360700

[56] Joseph CR Licklider. 1960. Man-computer symbiosis. *IRE transactions on human factors in electronics* 1 (1960), 4–11.

[57] Zhicong Lu, Mingming Fan, Yun Wang, Jian Zhao, Michelle Annett, and Daniel Wigdor. 2019. InkPlanner: Supporting Prewriting via Intelligent Visual Diagramming. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 277–287. doi:10.1109/TVCG.2018.2864887

[58] W. Mackay and D. Pagani. 1994. Video mosaic: laying out time in a physical space *(MULTIMEDIA '94)*. Association for Computing Machinery, New York, NY, USA, 165–172. doi:10.1145/192593.192646

[59] Nolwenn Maudet, Ghita Jalal, Philip Tchernavskij, Michel Beaudouin-Lafon, and Wendy E. Mackay. 2017. Beyond Grids: Interactive Graphical Substrates to Structure Digital Layout. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5053–5064. doi:10.1145/3025453.3025718

[60] Microsoft. 2024. Microsoft Word - Word Processing Software | Microsoft 365. https://www.microsoft.com/en-us/microsoft-365/word Accessed: 2024-09-12.

[61] Bonnie A Nardi. 1995. Studying context: A comparison of activity theory, situated action models, and distributed cognition. (1995).

[62] Don Norman. 2014. *Things that make us smart: Defending human attributes in the age of the machine*. Diversion Books.

[63] Sharon Oviatt. 1999. Ten myths of multimodal interaction. *Commun. ACM* 42, 11 (nov 1999), 74–81. doi:10.1145/319382.319398

[64] Amy Pavel, Dan B. Goldman, Björn Hartmann, and Maneesh Agrawala. 2015. SceneSkim: Searching and Browsing Movies Using Synchronized Captions, Scripts and Plot Summaries. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (Charlotte, NC, USA) *(UIST '15)*. Association for Computing Machinery, New York, NY, USA, 181–190. doi:10.1145/2807442.2807502

[65] Amy Pavel, Colorado Reed, Björn Hartmann, and Maneesh Agrawala. 2014. Video digests: a browsable, skimmable format for informational lecture videos. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 573–582. doi:10.1145/2642918.2647400

[66] Emilia Rosselli Del Turco and Peter Dalsgaard. 2023. "I wouldn't dare losing one": How music artists capture and manage ideas. In *Proceedings of*

the 15th Conference on Creativity and Cognition (Virtual Event, USA) *(C&C '23)*. Association for Computing Machinery, New York, NY, USA, 88–102. doi:10.1145/3591196.3593338

[67] Steve Rubin, Floraine Berthouzoz, Gautham Mysore, Wilmot Li, and Maneesh Agrawala. 2012. UnderScore: musical underlays for audio stories. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (Cambridge, Massachusetts, USA) *(UIST '12)*. Association for Computing Machinery, New York, NY, USA, 359–366. doi:10.1145/2380116.2380163

[68] Steve Rubin, Floraine Berthouzoz, Gautham J. Mysore, Wilmot Li, and Maneesh Agrawala. 2013. Content-based tools for editing audio stories. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology* (St. Andrews, Scotland, United Kingdom) *(UIST '13)*. Association for Computing Machinery, New York, NY, USA, 113–122. doi:10.1145/2501988.2501993

[69] Yong Rui, Thomas S Huang, and Sharad Mehrotra. 1999. Constructing table-of-content for videos. *Multimedia systems* 7 (1999), 359–368.

[70] Daniel M. Russell, Mark J. Stefik, Peter Pirolli, and Stuart K. Card. 1993. The cost structure of sensemaking. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands) *(CHI '93)*. Association for Computing Machinery, New York, NY, USA, 269–276. doi:10.1145/169059.169209

[71] Donald A Schön. 2017. *The reflective practitioner: How professionals think in action*. Routledge.

[72] Evan Schrier, Mira Dontcheva, Charles Jacobs, Geraldine Wade, and David Salesin. 2008. Adaptive layout for dynamically aggregated documents. In *Proceedings of the 13th International Conference on Intelligent User Interfaces* (Gran Canaria, Spain) *(IUI '08)*. Association for Computing Machinery, New York, NY, USA, 99–108. doi:10.1145/1378773.1378787

[73] Leixian Shen, Yizhi Zhang, Haidong Zhang, and Yun Wang. 2023. Data player: Automatic generation of data videos with narration-animation interplay. *IEEE Transactions on Visualization and Computer Graphics* (2023).

[74] 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 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1039, 19 pages. doi:10.1145/3613904.3642754

[75] Lucille Alice Suchman. 1987. *Plans and situated actions: The problem of human-machine communication*. Cambridge university press.

[76] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2024).

[77] Sangho Suh, Martinet Lee, Gracie Xia, et al. 2020. Coding strip: A pedagogical tool for teaching and learning programming concepts through comics. In *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 1–10.

[78] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. In *Symposium on User Interface Software and Technology (UIST)*. https://doi.org/10.1145/3586183.3606756

[79] Sangho Suh, Jian Zhao, and Edith Law. 2022. CodeToon: Story Ideation, Auto Comic Generation, and Structure Mapping for Code-Driven Storytelling. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 13, 16 pages. doi:10.1145/3526113.3545617

[80] Nicole Sultanum, Fanny Chevalier, Zoya Bylinskii, and Zhicheng Liu. 2021. Leveraging Text-Chart Links to Support Authoring of Data-Driven Articles with VizFlow. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 16, 17 pages. doi:10.1145/3411764.3445354

[81] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[82] Michael Terry, Chinmay Kulkarni, Martin Wattenberg, Lucas Dixon, and Meredith Ringel Morris. 2023. AI alignment in the design of interactive AI: Specification alignment, process alignment, and evaluation support. *arXiv preprint arXiv:2311.00710* (2023).

[83] Maryam Tohidi, William Buxton, Ronald Baecker, and Abigail Sellen. 2006. Getting the right design and the design right. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI '06)*. Association for Computing Machinery, New York, NY, USA, 1243–1252. doi:10.1145/1124772.1124960

[84] Anh Truong, Floraine Berthouzoz, Wilmot Li, and Maneesh Agrawala. 2016. Quickcut: An interactive tool for editing narrated video. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 497–507.

[85] Khai N. Truong, Gillian R. Hayes, and Gregory D. Abowd. 2006. Storyboarding: an empirical determination of best practices and effective guidelines. In *Proceedings of the 6th Conference on Designing Interactive Systems* (University Park, PA, USA) *(DIS '06)*. Association for Computing Machinery, New York, NY, USA, 12–21. doi:10.1145/1142405.1142410

[86] Theophanis Tsandilas, Catherine Letondal, and Wendy E. Mackay. 2009. Musink: composing music through augmented drawing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 819–828. doi:10.1145/1518701.1518827

[87] Barbara Tversky, Julie Bauer Morrison, and Mireille Betrancourt. 2002. Animation: can it facilitate? *International journal of human-computer studies* 57, 4 (2002), 247–262.

[88] Veed.io. 2024. Veed.io: Online Video Editing Platform. https://www.veed.io Accessed: 2024-12-08.

[89] Stephen Voida, Elizabeth D. Mynatt, and W. Keith Edwards. 2008. Re-framing the desktop interface around the activities of knowledge work. In *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology* (Monterey, CA, USA) *(UIST '08)*. Association for Computing Machinery, New York, NY, USA, 211–220. doi:10.1145/1449715.1449751

[90] April Yi Wang, Anant Mittal, Christopher Brooks, and Steve Oney. 2019. How Data Scientists Use Computational Notebooks for Real-Time Collaboration. *Proc. ACM Hum. Comput. Interact.* 3, CSCW (2019), 39:1–39:30. doi:10.1145/3359141

[91] Bryan Wang, Yuliang Li, Zhaoyang Lv, Haijun Xia, Yan Xu, and Raj Sodhi. 2024. LAVE: LLM-Powered Agent Assistance and Language Augmentation for Video Editing. *arXiv preprint arXiv:2402.10294* (2024).

[92] Bryan Wang, Yuliang Li, Zhaoyang Lv, Haijun Xia, Yan Xu, and Raj Sodhi. 2024. LAVE: LLM-Powered Agent Assistance and Language Augmentation for Video Editing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 699–714. doi:10.1145/3640543.3645143

[93] Fengjie Wang, Yanna Lin, Leni Yang, Haotian Li, Mingyang Gu, Min Zhu, and Huamin Qu. 2024. OutlineSpark: Igniting AI-powered Presentation Slides Creation from Computational Notebooks through Outlines. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 159, 16 pages. doi:10.1145/3613904.3642865

[94] Fengjie Wang, Xuye Liu, Oujing Liu, Ali Neshati, Tengfei Ma, Min Zhu, and Jian Zhao. 2023. Slide4N: Creating Presentation Slides from Computational Notebooks with Human-AI Collaboration. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 364, 18 pages. doi:10.1145/3544548.3580753

[95] Miao Wang, Guo-Wei Yang, Shi-Min Hu, Shing-Tung Yau, Ariel Shamir, et al. 2019. Write-a-video: computational video montage from themed text. *ACM Trans. Graph.* 38, 6 (2019), 177–1.

[96] Sitong Wang, Samia Menon, Tao Long, Keren Henderson, Dingzeyu Li, Kevin Crowston, Mark Hansen, Jeffrey V Nickerson, and Lydia B Chilton. 2023. Reelframer: Co-creating news reels on social media with generative ai. *arXiv preprint arXiv:2304.09653* (2023).

[97] Zijie J. Wang, Katie Dai, and W. Keith Edwards. 2022. StickyLand: Breaking the Linear Presentation of Computational Notebooks. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 269, 7 pages. doi:10.1145/3491101.3519653

[98] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[99] Steve Whittaker and Brian Amento. 2004. Semantic speech editing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vienna, Austria) *(CHI '04)*. Association for Computing Machinery, New York, NY, USA, 527–534. doi:10.1145/985692.985759

[100] 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.

[101] Haijun Xia. 2020. Crosspower: Bridging Graphics and Linguistics. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 722–734. doi:10.1145/3379337.3415845

[102] Haijun Xia. 2024. Redesigning the Information Space to Unleash the Power of AI. https://youtu.be/QSxKH8648WE [Accessed 31-03-2024].

[103] Haijun Xia, Jennifer Jacobs, and Maneesh Agrawala. 2020. Crosscast: adding visuals to audio travel podcasts. In *Proceedings of the 33rd annual ACM symposium on user interface software and technology*. 735–746.

[104] Haijun Xia, Jennifer Jacobs, and Maneesh Agrawala. 2020. Crosscast: Adding Visuals to Audio Travel Podcasts. In *UIST '20: The 33rd Annual ACM Symposium*

on User Interface Software and Technology, Virtual Event, USA, October 20-23, 2020. ACM, 735–746. doi:10.1145/3379337.3415882

[105] 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* (New York City, NY, USA) *(L@S '22)*. Association for Computing Machinery, New York, NY, USA, 163–174. doi:10.1145/3491140.3528279

[106] Haijun Xia, Nathalie Henry Riche, Fanny Chevalier, Bruno Rodrigues De Araújo, and Daniel Wigdor. 2018. DataInk: Direct and Creative Data-Oriented Drawing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*. ACM, 223. doi:10.1145/3173574.3173797

[107] Haijun Xia, Tony Wang, Aditya Gunturu, Peiling Jiang, William Duan, and Xiaoshuo Yao. 2023. CrossTalk: Intelligent Substrates for Language-Oriented Interaction in Video-Based Communication and Collaboration. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 60, 16 pages. doi:10.1145/3586183.3606773

[108] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. doi:10.1145/3544548.3581388

[109] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

[110] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

[111] 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.

[112] 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* (, San Francisco, CA, USA,) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 5, 30 pages. doi:10.1145/3586183.3606800

A Appendix

A.1 User Study Creation Outputs

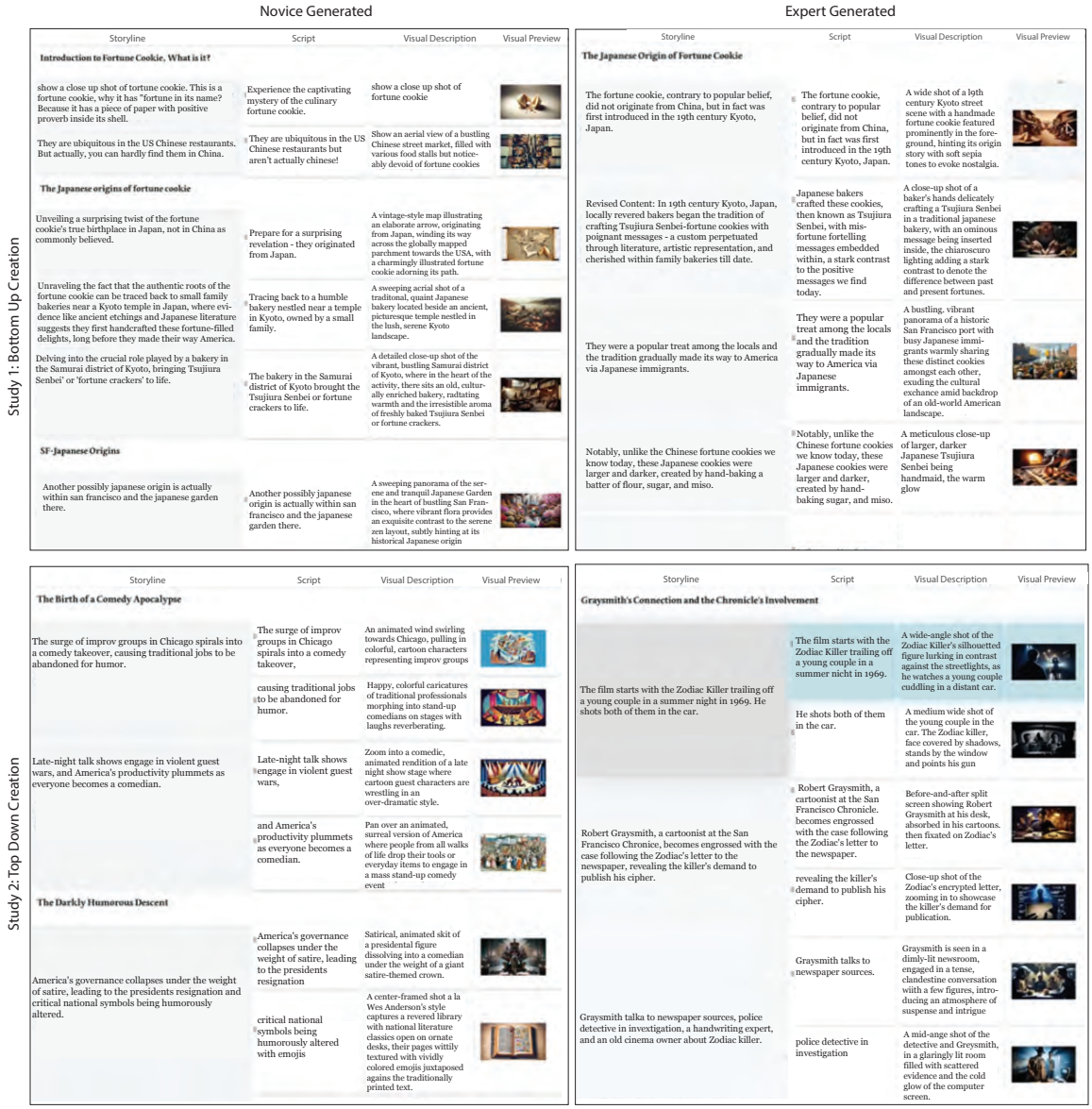


Figure 9: Samples of outputs created during our user study. The top two outputs were created in Study 1. The bottom two outputs were created during study 2. Outputs in the left column were generated by novices while outputs in the right column were generated by experts.

A.2 User Study Supplemental Data

Table 2: Expert Participants Demographic Data.

Participant	Video Creation Experience	Video Publishing Frequency	Video Creation Domain	Experienced with AI Video Creation?	Attended study 1	Attended study 2
E1	More than 5 years	Once every few months	Commercial video	Yes	Yes	Yes
E2	More than 5 years	Once every few months	Animation video, Film	Yes	Yes	Yes
E3	2-5 years	Once every few months	Film	No	Yes	Yes
E4	More than 5 years	Monthly	Vlog, Knowledge sharing video	Yes	Yes	No

Table 3: User Study Survey Results: Participants were asked a series of 5-point Likert-scale questions about their experience. They rated their sense of creative freedom and transparency when using the system as well as how the AI automations impacted their productivity and creativity. They also rated utility of the structures in helping them complete various tasks.

Category	Factor	Novices		Experts	
		Mean	SD	Mean	SD
Overall	Creative Freedom	4.67	0.52	4	0.82
	Transparency	4.33	0.52	3.5	1
Automations					
Synchronize Functions	Productivity	5	0	5	0
	Creativity	3.33	0.82	3.5	0.58
Generative Functions	Productivity	4.83	0.41	5	0
	Creativity	4.5	0.84	3.5	0.96
Structure Utility					
Canvas	Brainstorming	4	0.89	3.5	1.29
	Asset Organization	3.83	1.33	4	0.82
Narrative Editor	Brainstorming	3	1.27	4.25	0.96
	Narrative Development	3.83	1.83	4.5	0.58
Scene Planner	Brainstorming	4.16	0.98	2.75	0.96
	Narrative Developmen	4.5	1.22	4.25	0.5
	Scene Planning	5	0	5	0
Timeline Editor	Narrative Developmen	3.17	1.72	2.25	1.26
	Scene Planning	3.5	1.76	2.25	1.26
	Previewing Results	5	0	5	0
	Pacing Adjustment	4.17	1.33	5	0

A.3 Literature Analysis Results by Domain

Table 4: Compositional Structures in Different Content Creation Domains

Domain	Compositional Structures	Content Aspects	Function	Prior Works	Design Decisions Revolving Around Compositional Structures
Writing (Creative, Argumentative, Academic)	Canvas	Narrative	Material organization, Ideation, Notetaking	[40, 57]	"organize argument structures through synchronized text editing and visual programming" [112] "integrate multiple prewriting strategies into an iterative and flexible workflow" [57] "we focused our design on facilitating the transfer of information across applications while tracking its provenance" [40]
	Narrative Graph	Narrative	Narrative exploration and development	[24, 50, 57, 112]	
	Text Editor	Narrative, Spatial	Narrative development, Production	[40, 60]	
	Layout Editor	Spatial	Layout exploration and generation	[5, 59, 72]	
Podcast	Time-aligned Transcript	Narrative, Temporal	Transcript-based clip editing (Rough Cut)	[67, 68, 99]	"edit the transcript directly using standard word processing 'cut and paste' operations, which extract the corresponding time-aligned speech" [99]
	Timeline	Temporal	Timeline-based clip editing (Fine Cut)	[4, 28]	
Music	Canvas	Narrative	Material organization, Ideation, Notetaking	[20, 38, 66]	"Composers create their own individual ad hoc strategies for expressing ideas, and often move back and forth between multiple representations " [39] "(Canvas) offering composers the freedom to arrange scores and musical fragments spatially, adapting the layout to the specific task at hand. " [20]
	Tone-Network	Congruent	Chord composition	[37]	
	Staff	Congruent, Temporal	Notation-based composition	[38, 86]	
	Timeline	Congruent, Temporal	Production	[1, 8, 9]	
Interactive Text-Visual Narrative (Interactive Article and Comics)	Canvas	Spatial, Narrative	Material organization, Ideation, Sketching	[51, 97]	"making the correspondence explicit and consistent is essential when presenting multiple representations… the mapping between code, story, and comic should be clear" [77] "a section of a narrative and its corresponding visuals will be organized within one block throughout the entire design process to enable flexible prototyping ... while ensuring their correspondence" [18]
	Panel-based Editor	Narrative, Spatial	Narrative and layout development	[10, 49, 77, 79]	
	Section-based Editor	Narrative, Spatial	Narrative and layout development	[18, 80]	
	Notebook-based Editor	Narrative, Spatial	Narrative development, Visual creation	[49, 93, 94]	
	Enhanced Text Editor	Narrative, Spatial	Narrative development, Interaction creation	[21, 25, 42]	
Video	Media Gallery	-	Media organization	[6, 92]	"Establishing elastic and customized mappings between animation and performance to enable graphic elements to adapt to real-time speech and gestures to achieve synchronization and expressive presentation effects." [19] "enabling synchronized browsing of the captions, script and summary for easy access to details or context at any point in the film" [64]
	Storyboard	Narrative, Spatial	Narrative development, Asset arrangement.	[41, 43, 44, 58]	
	Time-aligned Transcript	Narrative, Temporal	Transcript-based clip editing (Rought Cut)	[28, 73, 84, 104]	
	Timeline	Congruent, Temporal	Timeline-based clip editing (Fine Cut)	[6, 28, 92]	

A.4 Implementation of AI Integration in VideOrigami: Generation Details and Prompts

Table 5: AI Functionalities within Canvas

AI Feature	Context	Generation Details	Prompt Used for Generation
Generate notes based on the prompt	Parent Asset node	Parse and query document using OpenAI Assistant API	*{prompt}. make your response in the most concise way possible.*
Generate description/caption for images	Image in the Asset node	Query image using OpenAI API GPT4-visual-preview	*Describe the visual scene in the image to a filmmaker in a concise way. Consider shot type and cinematic style. Make your response as short and concise as possible. Only use 1 sentence.*
Regenerate image based on revised prompt	Image in the Asset node	Generate image with prompt using OpenAI API Dall-E-3	*Generate an image based on the prompt exactly. do not change or revised prompt for generation: {prompt}*

Table 6: AI Functionalities within Narrative Editor

AI Feature	Context	Generation Details	Prompt Used for Generation
Generate talking points within a section	Section heading	Generate text using OpenAI GPT-4 chat completions API → parse the coded response	*you are a video creator for a video about for- tune cookie origin. You need to come up with compelling narratives and visuals for the video planning. You task is generate several talking points within the given section. The talking points should present a narrative that fits into the section. The talking points should commu- nicate what you want to deliver in each scene. Each talking point should be one sentence long. Talking points should have specific details. sec- tion: {content} response with a string that contains 2-4 concise and informative talking points. The talking points should flow logically. They should not repeat each other. separate each talking point with ###. don't index them, don't add quotes or prefix.*
Generate talking points / sections	Assets, notes in canvas and existing content in the editor	Generate text in context of document using OpenAI Assistant threads API	*Consider the video creation prompt: {query}. The video should be according to the content in the file. Give me the possible section headings and talking points that would be in such a video and that would form a cohesive narrative. Each heading should be a short sentence. Each talk- ing point should be one sentence long. Talking points should have specific details. The talking points should flow logically. They should not repeat each other. Give me 2 section headings and 2-3 concise and informative talking points within each section. Give me the response as one string. Each section heading should be pre- fixed by %% and each talking point should be prefixed by ##. First, give me the section heading and talking points of the first section. Then give me the section heading and talking points of the second section. Make your response in the most concise way possible. Don't include sources. Ex- ample response: %%Section 1##Talking point 1##Talking point 2%%Section 2##Talking point 3##Talking point 4*

Table 7: AI Functionalities within Scene Planner Grid

AI Feature	Context	Generation Details	Propmt Used for Generation
Refine script	Cells in the same row and column	Generate text using OpenAI GPT-4 chat completions API	*You are video producer. Your task is to refine text be a voice-over. Make it as engaging, direct and concise as possible. text to refine: **{content}***
Suggest split of script based on Scene	*null*	Generate text using OpenAI GPT-4 chat completions API → Parse the coded response	*You are a professional video producer, trying to segment the script. Separate the text narrative delimited by triple backticks into segments. Different segments can be the result of different accompanying visuals or changing in subject. Consider breaking sentences into multiple segments if the sentence can be represented by different visuals. Format the response as a list of strings (where the strings are substrings of the given text narrative). Make your response as short and concise as possible. text description: **{script}***
Visual description generation	Visual style, script, and visual descriptions of scenes in the same storyline paragraph	Generate text using OpenAI GPT-4 chat completions API	*You are a professional video producer, trying to couple the visual with the script. Describe the visual you will use based on the list of text narrative segments and accompanying visuals delimited by triple backticks and intended visual style. The following is a list of dictionaries with 2 keys: visual description and script segment. List of segments and visual descriptions: **{script}** Overall visual style: **{style}** (ignore if N/A) You want to describe a static visual scene for the list item where the visual description field says "PROVIDE". Consider the shot type and visual style. Format your response as a string that describes a visual scene that can be pictured with one image. Make your response as short and concise as possible. Only use 1 sentence.*
Visual preview generation	Script, generation style, all visual descriptions, storyline paragraph	If no visual description, generates visual description → generate image using OpenAI image generation with Dall-E-3 → Parse the coded response	***{description}**, in **{style}** aesthetic.*

Table 8: AI Functionalities within Timeline Editor

AI Feature	Context	Generation Details	Propmt Used for Generation
Align imported audio file with the existing time segments in the audio track	*null*	Generate transcription of the audio with timestamps using OpenAI Whisper-1 → segment the transcription text to align with the script column in the grid → parse coded response	*Segment the audio transcription according to the original script segments. Original script segments: **{original segments}** transcription: **{transcript}** Return the transcription with ### in the places where you plan to split the transcription into segments.*

Table 9: AI Functionalities across Structures

Structures	AI Feature	Context	Generation Details	Propmt Used for Generation
N ⇒ C	Generate notes based on talking points	Parent Asset and all narrative content within a section	Parse and query document using OpenAI Assistant API	*Getting relevant content for {talking point}*
N|S ⇒ C	Finding relevant notes while writing script	narrative content being edited	Create embeddings for notes and talking point using OpenAI API text-embeddings-3-small → Get notes whose embeddings are less than 0.93 distance apart from the talking point embedding	
C ⇒ N	Form talking points with note	All content in the editor	Generate text with OpenAI GPT-4 chat completions API	*Your job is to revise the current content with the note, make it fit into the existing narrative. the current content is part of the exiting narrative. You need to understand where current content is, and how to make it more solid with the note and how the revised version can smoothly fit into the narrative flow. current content:* ***{current content}****; note:* ***{note content}*** *existing narrative:* ***{talking points}****. Your generated content should be as direct and concise as possible. one sentence.*
N ⇒ S	Populate storyline column to the relevant script/visual columns	*null*	Generate text with OpenAI GPT-4 chat completions API with few-shot prompting *	*Your task is to segment the storyline content into voicee over and visual description. You will be provided with the sandbox content. response a dictionary of the segmented result." sandbox content:* ***{content}*** *only use the content provided, don't add new content!*
T ⇒ S	Fine-tune script by adjusting the time segments	Existing script in Scene Planner	Generate text using OpenAI GPT-4 chat completions with few-shot prompting → parse coded response	*You are a wordsmith. Make the text* ***{length}*** *words* ***{shorter | longer}****. Do not change the meaning of the sentence, but you can add or remove words. The output sentence should be meaningful and cohesive. Text:* ***{content}*** *Give the original text with annotations. Put ### around the words that you added to the original sentence. Put removed from the original sentence. Respond with only the original text with annotations. Do NOT prefix the response with anything.*
S ⇒ T	Generate audio voice over based on script	*null*	Generate audio voice over based on script & Generate audio file using OpenAI Text-To-Speech API -> generate transcrip- tion of the new audio file to get timestamps for words using OpenAI Whisper-1	