AI-Instruments: Embodying Prompts as Instruments to Abstract & Reflect Graphical Interface Commands as General-Purpose Tools

Nathalie Riche
Microsoft Research
Redmond, Washington, USA
nath@microsoft.com

Anna Offenwanger
Microsoft Research
Redmond, Washington, USA
CNRS, Inria, LISN
Université Paris-Saclay
Orsay, France
anna.offenwanger@gmail.com

Frederic Gmeiner
Microsoft Research
Redmond, Washington, USA
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
gmeiner@cmu.edu

David Brown
Microsoft Research
Redmond, Washington, USA
dabrown@microsoft.com

Hugo Romat
Microsoft
Seattle, Washington, USA
romathugo@microsoft.com

Michel Pahud
Microsoft Research
Redmond, Washington, USA
mpahud@microsoft.com

Nicolai Marquardt
Microsoft Research
Redmond, Washington, USA
nicmarquardt@microsoft.com

Kori Inkpen
Microsoft Research
Redmond, Washington, USA
kori@microsoft.com

Ken Hinckley
Microsoft Research
Redmond, Washington, USA
kenh@microsoft.com

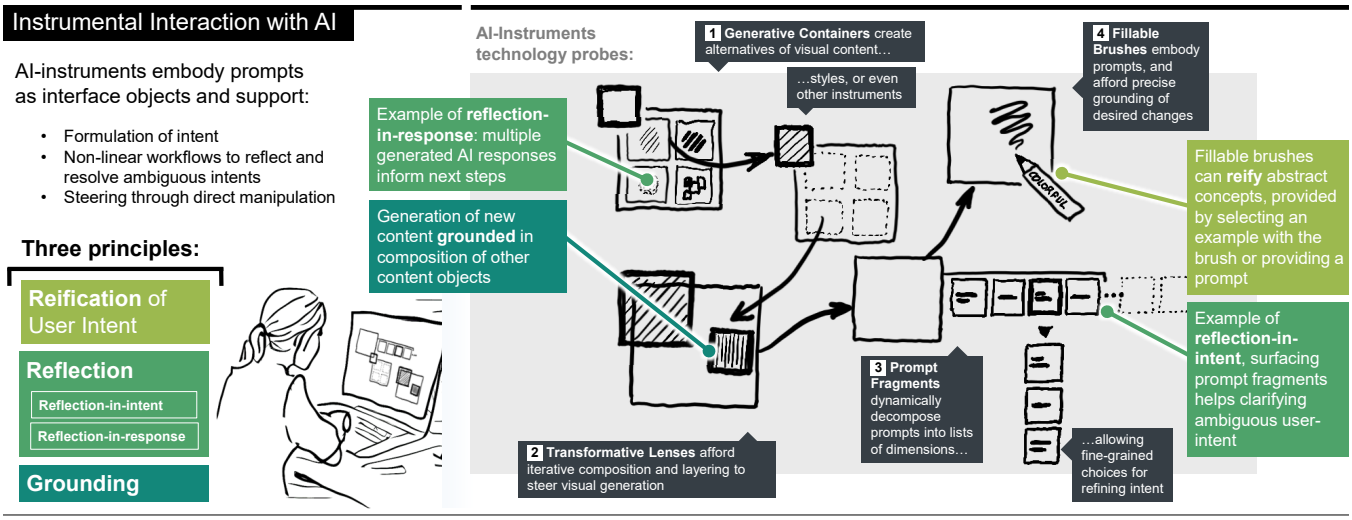


Figure 1: AI-Instruments embody prompts as interface objects, informed by three principles: reification of user intent, reflection, and grounding (left). Visual overview of four technology probes of AI-Instruments – generative containers, transformative lenses, prompt fragments, and fillable brushes (right).

Abstract

Chat-based prompts respond with verbose linear-sequential texts, making it difficult to explore and refine ambiguous intents, back up and reinterpret, or shift directions in creative AI-assisted design work. *AI-Instruments* instead embody "prompts" as interface objects via three key principles: (1) *Reification* of user-intent as reusable direct-manipulation instruments; (2) *Reflection* of multiple interpretations of ambiguous user-intents (*Reflection-in-intent*) as well as the range of AI-model responses (*Reflection-in-response*) to inform design "moves" towards a desired result; and (3) *Grounding* to instantiate an instrument from an example, result, or extrapolation directly from another instrument. Further, AI-Instruments leverage LLM's to suggest, vary, and refine new instruments, enabling a system that goes beyond hard-coded functionality by generating its own instrumental controls from content. We demonstrate four

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714259

technology probes, applied to image generation, and qualitative insights from twelve participants, showing how AI-Instruments address challenges of intent formulation, steering via direct manipulation, and non-linear iterative workflows to reflect and resolve ambiguous intents.

CCS Concepts

• **Human-centered computing → Interaction paradigms**.

Keywords

instrumental interaction, generative AI interfaces

ACM Reference Format:

Nathalie Riche, Anna Offenwanger, Frederic Gmeiner, David Brown, Hugo Romat, Michel Pahud, Nicolai Marquardt, Kori Inkpen, and Ken Hinckley. 2025. AI-Instruments: Embodying Prompts as Instruments to Abstract & Reflect Graphical Interface Commands as General-Purpose Tools. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3706598.3714259

1 Introduction

Despite the immense promise of generative AI, it remains challenging for people to express and refine their true intents via multiple rounds of textual chat-prompts [48, 78], as well as to pursue multiple-alternative paths forward within a linear conversational metaphor. Users face numerous difficulties (e.g., [64]): articulating their intent in a few words of written text (intent formulation); correctly expressing sufficient detail to express, refine, and re-formulate their true intent (intent disambiguation); iterating over the model's response to approach a desired outcome (steering); and navigating higher-order challenges of interaction with AI, such as discovering what one can do with AI—or even what one "actually" wants to achieve (intent resolution)—within the linear sequential limitations of chat-based exchanges (interaction workflow).

Although existing work in human-computer interaction and artificial intelligence (HCI+AI) addresses some aspects of these challenges in piecemeal fashion through novel interaction techniques with generative AI (e.g. [17, 50]), we argue here that *AI-Instruments* offer a novel approach to gain traction on many aspects of these challenges by appropriating and re-casting the principles of *instrumental interaction* [5, 7] to the modern context of generative HCI+AI user experiences. Quoting Beaudoin-Lafon et al. [6], the value of proposing such interaction model is to *"change-oriented perspective by providing HCI researchers with conceptual tools for analyzing technologies in use or exploring novel future solutions"*. Triangulating theory, artifact, and empirical evaluation has strong benefits for advancing HCI research [47].

Instrumental interaction offers a particularly compelling concept from the HCI literature to revisit in the context of generative AI because it offers principled interaction dynamics about how software functionalities (*"commands"*) combine with content (the *"objects"* those commands act upon). While in the past these dynamics had to be hand-designed and hand-coded for specific object types and application settings, the advent of generative AI makes it plausible that the polymorphic nature of high-level commands and flexible content representations will unleash exciting new possibilities for HCI+AI graphical user interfaces.

In particular, our approach *embodies AI prompts as graphical interface objects* and adapts the instrumental interaction model for Generative AI by considering the following three principles:

(1) ***Reification of user intent*** into instruments: turning user-intent from varied abstractions and granularity levels into one or more reusable graphical interface object(s);
(2) ***Reflection***: the consideration of multiple alternatives that reflect [61] both ambiguous intents as expressed by the user (*reflection-in-intent*), and ambiguous interpretation of AI responses (*reflection-in-response*), to steer content generation towards a satisfactory result; and finally
(3) ***Grounding***: instantiating an instrument from a specific scope of selected content, from an example result, or even from another instrument.

Via a technology probe [34] that implements four complementary examples of AI-instruments, we illustrate how they can ameliorate many design challenges plaguing today's linear-chat-based AI interfaces: intent formulation, prompt engineering, direct manipulation and steering, non-linear iterative workflows, and intent resolution. We also present initial reactions of 12 participants who tried our AI-instruments, yielding qualitative insights on the value and limitations of our AI-instruments interaction model, as compared to conversational prompting.

Designed through the lens of the three principles, we built a set of technology probes focused on image generation. The goal of these four exemplar AI-Instruments is to demonstrate the new interaction capabilities and affordances of our model: (1) *Fragments* decompose gen-AI prompts into reified reconfigurable objects, affording reflection-on-intent on the latent prompt structure, and grounding generation by dragging fragments from one object to another. (2) *Transformative Lenses* generate new content grounded in one or more content elements, which allows flexible recomposition of scenes and (if desired) continuous updates of the result. (3) *Generative Containers* create multiple alternatives of images, text, and even instruments or fragments. (4) *Fillable Brushes* encapsulate a prompt, filled by selecting example content with the brush (or by directly typing the prompt for a new action). Using the instruments in synergy—where outputs from one instrument form input for the next, or even using instruments to create new meta-instruments—affords expressive degrees-of-freedom for fine-grained steering of generative AI.

In summary, our high-level contributions include the following:

- Extend the classic instrumental interaction model [5] to generative AI, emphasizing three driving principles: reification of user intent, reflection, and grounding;
- Demonstrate four AI-instruments via technology probes, showing how these driving principles manifest in their design and implementation;
- Provide initial reactions from 12 users when shifting from a linear-chat interaction paradigm to direct manipulation through AI-instruments, showing that it can address a number of human-AI interaction challenges.

In the following sections we discuss related techniques across the HCI, Human-AI interaction, and design literature. This is followed

by an Example Walkthrough of our AI-instruments, a wider discussion of Instrumental Interaction with AI, and further details of our Four Exemplar AI-Instruments: Fragments, Transformative Lenses, Generative Containers, and Fillable Brushes. We then present a qualitative Study of these AI-Instruments in comparison with textual prompting, and finally close with a Discussion and Future Work.

2 Related Work

We first discuss the state of human-AI interaction, articulating it around five core challenges. Then, we motivate the need for a more general interaction model and point to research in design and creativity grounding two principles we introduce.

2.1 Human-AI Interaction

Recent work explores the difficulties users face when interacting with generative AI via prompting [48, 54, 64, 78]. Earlier research identified barriers that arise (for example) in end-user programming [42] and, more generally, bridging the gulf of execution and the gulf of evaluation [51]. We organize emerging research for interaction with generative AI under five core Challenges (C1-C5) faced by users, and discuss later in the paper how our interaction model addresses each.

(C1) Intent formulation via prompting, solely using natural language, can be challenging when the outcome is hard to describe in words. Users may lack the vocabulary to describe visual styles, or the high-level impressions they seek to achieve. Researchers studied thousands of prompts to generate images [45] to develop guidelines for prompting and parameter selections. They coupled prompting with images to offer richer multimodal intent formulation. PromptCharm [72] leveraged a large image database to help users find the right style of images and incorporated interactive techniques – such as linking a prompt fragment to the corresponding part of the generated image, to provided richer solutions for users to formulate their intent. Similarly, DesignPrompt [54] affords expressive multi-modal prompt construction. Such research seeking to expand the modalities we have to communicate with models beyond text input is particularly important [44] for multimodal outputs such as generated videos [70], 3D objects [55], and virtual worlds [59].

(C2) Intent disambiguation is the skill of describing one's intent with enough specific detail for AI to produce the intended result. Much past work on *prompt engineering* across several fields of research tackles this challenge, with research probes of this issue [78] suggesting templates and guidelines [9] for users to provide the information they might have difficulty thinking about upfront. Beyond prompt engineering, the HCI community explores different representations to facilitate communicating context to the system. For example, Graphologue [37] represents a prompt as an interactive node-link diagram that users can expand and complete to incrementally add context to their intent. Such work also addresses the ambiguity of natural language by enabling users to unpack certain parts of their intent and disambiguate them by adding more information. Promptify [10] organizes generated content on a canvas based on a person's preferences and suggests alternatives – leading to an iterative loop with the user refining, selecting, and discarding alternatives of prompts and content.

(C3) Intent resolution is the challenge users confront to determine what outcomes may or may not match their original intent. Difficulties here may stem from an ambiguity of intent in the users' mind (e.g. a user might realize *"I am not even sure what exact outcome I want"*). This problem, as a well-known attribute of challenging creative design work [12, 28, 61], is certainly not unique to AI but may be exacerbated by the relative novelty, black-box nature, and rapidly accelerating capabilities of modern AI models [11]. However, further difficulties may arise from people's lack of knowledge of what an AI model can or cannot do. Here, approaches from graphic design may help users explore possible outcomes, such as CreativeConnect [16], which extracts keywords, and text descriptions from a set of reference images and facilitate recombination and reuse. Other work shows the possibilities of what users can ask via prompt-space exploration [2], or through interfaces that reveal what results a user can generate [65].

(C4) Steering the result of generative AI to get closer to either what the user initially imagined or to an unforeseen result assessed as satisfactory is a fundamental human-AI interaction mechanism. The topic has been studied for multiple decades in multiple field and referred to as human-in-the-loop [73] and mixed-initiative interfaces [32]. Within the context of generative AI, research on the topic has centered on human-AI co-creation [20]. Researchers developed human-AI co-creation interfaces for specific activities such as drawing [52], crafting images [17] and writing stories [18, 79]. These interfaces either surface generative AI capabilities as graphical interface elements such as a button to generate a character for a story [79], or propose custom graphical widgets to specify constraints or parameters of the content to be generated by the model such as an interactive line chart depicting the narrative arc of the story [18].

Recent research has begun to explore more generic interaction solutions to the prompting chat-based experiences incorporated in most mainstream products today. Steering content generation in conversational prompting amounts to a linear trial-and-error process, in which users type a prompt, and then evaluate its result. They then must either rerun the same prompt to get a new result (since generative AI is non-deterministic); or edit the prompt to get an iteration over the prior result. By building upon principles of *direct manipulation*, DirectGPT [50] offers an early glimpse of an alternative interaction human-AI co-creation paradigm based on the principles of direct manipulation and surfaced to users with graphical widgets (e.g. buttons) that might generalize to a wider range of outputs and applications. Our research extends this ambition via instrumental interaction [5], yielding a novel interaction model that can provide the community with both evaluative (assessing novel interaction techniques) and generative (inspiring the design of novel interaction techniques) power.

(C5) Interaction workflow models based on conversation with generative AI are inherently linear. Research started to investigate non-linear interaction workflows with generative AI. In particular, DeckFlow [19] relies on mood board type interaction and also

breaks the silo of different models. Sensecape [66] and Graphologue [37] leverage additional non-linear metaphors to enable people to perform non-linear interactions with AI. These systems focus on a specific metaphor for conversation with LLMs, laying out prompts and responses as a graph in a canvas. Graph structures can also function as an intermediary representation facilitating prompt steering, by breaking down text prompts into hierarchical structures of granular elements [76]. Similarly, tree structures enable traversing alternative representations of generated content, where sub-nodes represent distinct visual aspects across the latent space [71]. Our intent is to identify general principles that afford direct manipulation for decomposing and (re)composing objects at multiple levels of granularity for different tasks and contexts.

The interaction model we propose provides a generic solution to address these 5 challenges by building upon the instrumental interaction model and apply it to the context of building interactive applications leveraging generative AI capabilities.

2.2 Interaction Models in the Era of AI

Despite tremendous advances in technology and the promise of Artificial General Intelligence [11], mainstreams interfaces today feature a chat-based interface with AI reminiscent of command-line human-computer interaction paradigm of the 1960s. While the use of natural language does remove barriers of adoption for the general public, many of the limitations of communicating instructions in a linear and sequential manner by typing, later addressed by Graphical User Interfaces, pertain.

Over the years, the HCI community has produced knowledge on human-computer interaction [31], devised principles and theories for improving interaction [25, 33], and proposed multiple interaction models [5, 36] for building the next generation of interfaces. These models are generally grounded in the emerging interfaces and techniques of the time, surfacing key principles governing them and desirable properties when humans interact with them. The goal of these models is to inform and assess the design of the next generation of interfaces. Our work has the same ambition: *informing and guiding the design of interfaces leveraging generative AI.* While numerous recent work centered on advancing specific use cases and application areas – seeking to identify and leverage the value of generative AI – few researchers relate to existing theory and models, or proposing *new theories and models* in this era of AI. Perhaps the closest effort is the Cells, Generators, and Lenses model proposed by Kim et al. [39] which proposes a design framework for helping designers identify and reflect on basic building blocks needed for interfaces leveraging AI. Our research is complementary to this effort, seeking to identify interaction principles that afford direct manipulation of these building blocks.

Our work seeks to build upon and extend the instrumental interaction model to the design of generative AI interfaces. The instrumental interaction model [5] directly builds upon direct manipulation and generalizes the use of instruments to mediate between user and objects of interests (e.g. content). It describes a large range of interaction techniques that were not captured in WIMP and direct manipulation such as lenses or tangible interactions. Recent work attempted to leverage the instrumental interaction model to design novel interactions with AI. For example, Yen and Zhao [77] used reification to turn prior conversations with AI into graphical objects or Memolets, that users can interact with. Our work propose a more general adaptation of this model to content generation with AI and expands it with additional principles of reflection and grounding. The resulting model we propose falls into generative theories of interaction [6], aiming at inspiring and informing the design of novel techniques.

2.3 Content Generation and Creativity Support

Several key insights from the design and creativity support literature [62] motivate our principles of reflection and grounding.

Design and creativity processes embrace ambiguity of low-fidelity prototypes [12] and rapid cycles of idea generation and evaluations [26, 28, 67] to enable people to explore many design alternatives, reflect on their possibilities through the action of sketching and building, and iterate on the most promising ones [61]. Researchers have also described these processes as sequences of divergent thinking followed by convergent thinking [21]. As fundamental working-patterns that people exhibit in challenging content creation and design tasks, there is good reason to believe such processes should persist and be supported by tools for creative content generation with generative AI. Such tools should help users rapidly investigate alternatives ideas in fluid, non-linear manner (e.g. exploration) and support the rapid iteration of the most promising content (e.g. steering). Direct manipulation and instrumental interaction models offer a compelling point of departure, affording chunking and phrasing [14] of complex generative-AI interactions for exploration and steering.

As hinted above, our principle of reflection builds on Schön's notion of *reflection-in-action* [61]—where the externalized materials of design "speak to" the designer to help them reflect-*on*-action as to the next design "move" to make within an ambiguous space of many possible ideas. In the context of generative AI, this principle of *reflection* conveys the notion that instruments should reflect the design space of user intent—as well as the wide potential space of generated results—to help users make informed decisions ("moves") as they iterate towards a desired (AI-assisted) outcome. A related concept to reflection and idea incubation is the process of gathering inspirational materials in moodboards [12, 15]. Such design practices help identify concepts and themes, especially when these are hard to articulate in words, or isolate from one another other [23]. We refer to this activity in our principle of grounding, to convey the idea that instruments can extract specific aspects from a set of materials, and to then apply them to different content.

An aspiration for an interaction model geared on content generation is to afford power to their users [43]. In particular, vertical movement (moving up and down the abstraction ladder) afforded by natural language input of LLMs; and horizontal movement (composing tools and workflows) afforded by combining instruments together offer promising avenues for AI-instruments.

3 Example Walkthrough

Let us take the example of Emma, who is seeking to illustrate a social media post to express the serenity she feels when she spends time outdoors (Figure 2). She starts from leveraging generative AI to generate a bird. The art style is not quite satisfying but she

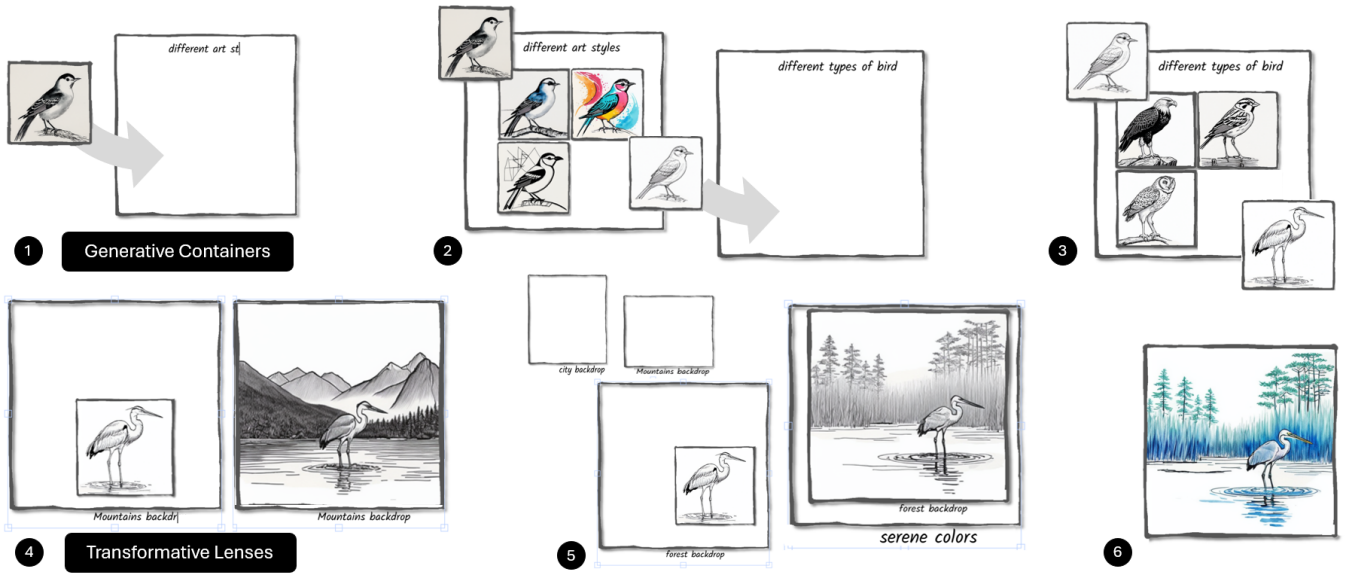


Figure 2: Sequence of interactions to explore ideas with generative containers and lens probes: When dragging an image into a container (1), variations are created based on *style* (2). When selecting one of these images and dragging it into another container with the prompt "different types of bird", variations of different kinds of birds are generated in a consistent art style (3). A transformative lens around one of the earlier images generates a landscape around the bird through inpainting (4), and allows more complex composition of content (5, 6).

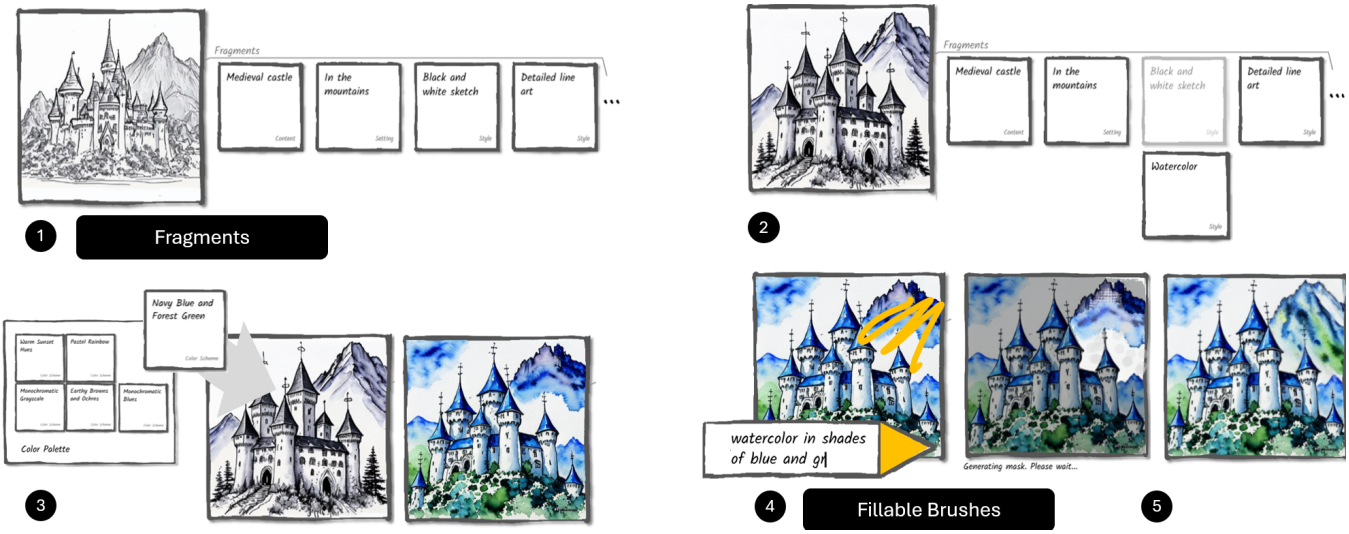


Figure 3: Sequence of interactions to steer image generation with fragments and brushes probes: Prompt fragments are generated for an existing image and show dimensions of the image to manipulate (1). A person can modify any of these fragments and a new image is generated (2). Containers can generate variations of fragments, which are then used to modify the image (3). Fillable Brushes (pen-like instruments) are used to modify the image of a castle, changing the art rendering style and color where the brush painted over the image, based on the prompt that was 'filled' into the pen (4, 5).

is not sure what the model is capable of. She selects a generative container from a panel of AI-instruments available to her (Figure 2.1 and 2) and explores different art styles. She finds a simple drawing style she likes, and creates a second generative container to explore other types of birds in the same style (Figure 2.3). She settles on a heron, and moves on composing a more interesting illustration. Since she has an idea of the general composition she wants, she opts for a transformative lens, a second AI-instrument enabling her

to position her central character, the bird, in the frame (Figure 2.4). She creates multiple lenses to try multiple backdrops and settles on the forest one (Figure 2.5). As lenses can be layered, she creates a color style one, and layers it on top of the forest backdrop, resulting in an illustration she finds suitable for her post (Figure 2.6).

Two days later, Emma seeks to illustrate her school presentation on medieval castles (Figure 3). She starts from a drawing generated by AI. She taps-and-holds to expand the fragments the model used to generate the image (Figure 3.1). By tapping on different fragments, she gets to try different variations, such as redrawing the castle as watercolor style (Figure 3.2). As she wants to add color to the illustration, Emma retrieves the palette where she saved multiple fragments related to colors she thought worked great in the past (Figure 3.3), and drags one onto the image. She does like the colors but notices a large white space in the back. She selects a fillable brush, an AI-instrument that lets her directly scrub over the specific portions of the image that she wants to revise or refine (Figure 3.4). After she types the outcome she wants and brushes the region, the system generates a mask and applies changes locally (Figure 3.5). Emma is now satisfied with her illustration.

4 Instrumental Interaction with AI

Beaudoin-Lafon defines instruments as: *"a mediator or two-way transducer between the user and domain objects."* we expand this definition to AI-instruments: *"an AI-powered mediator or two-way transducer between the user and domain objects."* We describe below the three principles of our proposed model revision: reification of user intent, reflection and grounding. Note that these principles are tightly interconnected and, while differing in certain aspects from the original model also share a lot of similarities. We discuss differences in more depth in Discussion.

4.1 Reification of User Intent

Most pre-AI interfaces offer a finite set of functionalities, established at their design by software architects and developer. User experience designers craft a set of graphical interface components and interactions for each functionality to enable users to invoke a finite set of commands through this GUI. Today, LLMs can interpret requests from users in natural language and turn them into the execution of a specific command, or a sequence of commands, unbounding functionalities from a limited set of GUI components. With this major shift in interface design, we propose the reification of **user intent**, rather than **commands**.

Reification turns both input and output of generative AI into graphical elements that can be directly manipulated and thus reused by users. In contrast to chat-based interfaces consisting of sequences of [input+output] in which users can require to rephrase the input to iterate, reifying input and output enables users to articulate phrases of interaction [14] and afford direct manipulation techniques such as lasso selections to specify scopes of intent (Figure 4 (1-3)). In section 5, we demonstrate how this instrumental model can leverage the full range of direct manipulation techniques the community developed such as magic lenses [8] and attribute objects [74], turning them into AI-instruments encapsulating user intent.

A key capability of generative AI models is their inherent ability to deal with the **degree of abstraction** of user intent. It offers unparalleled flexibility as users can express high-level or low-level intent. Examples in the literature leverage the high degree of abstraction for content generation. For example, Talebrush [18] enables users to control the narrative arc of a story (where tension is in the story), which has many implications on the writing itself from adding or sequencing events differently in the story to subtly rewording the language. Expressing high-level intents is a powerful ability, enabling people to shape content in ways that potentially lead to serendipitous discovery of alternative (potentially better) results. However, users face multiple challenges when results are unsatisfactory, understanding how to resolve ambiguity of their intent (C2) or thinking more crisply of the desired outcome (C3). These challenges often require users to lower the degree of abstraction of their intent. On the contrary, expressing intents with a low degree of abstraction lowers the chance to make serendipitous discoveries, as well as get into a class of unwanted model results, making it frustrating for users to steer content generation towards more major changes (C4) or conduct exploratory workflows (C5). These challenges often require users to increase the degree of abstraction of their intent. Figuring out how to navigate degrees of abstraction is a challenge in itself. Users may struggle turning an idea into a set of concrete changes or, conversely, articulate the overarching goal motivating specific changes. Users can leverage AI-instruments themselves to navigate the degree of abstraction of an intent, for example, by using a Generative Container to provide more concrete (resp. abstract) Fragments given one of high-degree (resp. low-degree) of abstraction (Figure 4 (4-5)).

4.2 Reflection

Seminal research demonstrated that it is critical to explore alternative designs early and throughout the whole process [49, 69]. It is particularly critical when working with AI because of its "black box" nature [4, 30], i.e. the inherent difficulty for users to understand how these models work, and the non-deterministic nature of their outputs. To capture this aspect, we borrow the term **reflection** from the design literature and introduce it as a principle for AI-instruments.

We define reflection as the ability to help users reflect on their possibly ambiguous intent (reflection-in-intent) as well as the ambiguous interpretation made by AI (reflection-in-response), and thus offer the ability to users to steer the content generation towards a satisfying result.

Reflection-in-intent is the ability of AI-instruments to surface multiple facets of their intent to users. For example, fragmenting intent into pieces reveals a particular chunking [14]. Working with fragments (Figure 5) may help users refine their intent (1), pivot on a specific aspect (2) or iterate by adding novel aspects (3).

Reflection-in-response is the ability of AI-instruments to offer multiple results of the content generation, while also helping people explore the space of possibilities (Figure 5), addressing (C3).Reflection-in-response can vary on the type and range of alternatives provided by employing diverse strategies: using model parameters such as its temperature, generating variations of the input, or asking the model to use different context of interpretation.

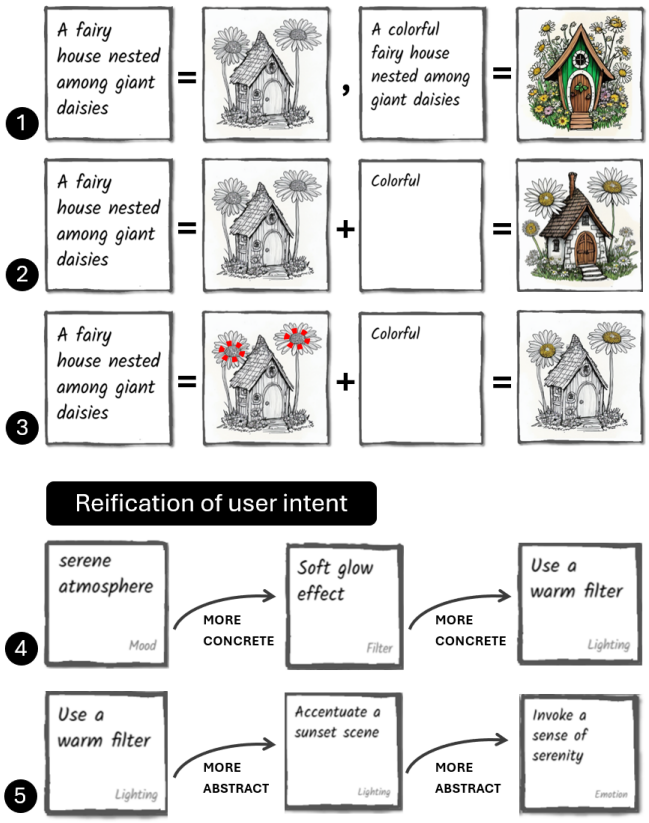


Figure 4: In the chat-based interaction model, interactions consists of a linear sequence of input+output pairs and steering is done by modifying the input (1). Reification enables articulating interactions into phrases for example by reusing the output of the prior input (2). It also affords direct manipulation techniques such as for lasso selection (in red) to specify the scope of the input (3). Reification of user intent enables users to reflect on their intent and navigate dimensions such as its degree of abstraction, using other instruments to make it more concrete (4) or abstract (5) for example.

4.3 Grounding

The principle of grounding refers to the ability for users to ground instruments from examples of desired outcomes or other instruments. It may be difficult to find the right vocabulary to describe particular aspects of content, especially for images. Instruments leverage AI segmentation to (1) enable users to refer to elements of an example in generic terms, and (2) extract specific aspects of the content (e.g. style) by selection, storing the result for later (Figure 6). This builds on the notion of *Variations*, *Parameter Spectrums*, and *Side Views* [34, 68], but in a way that leverages the principles of interactive instruments [5, 7] as well as the open-ended possibilities of generative AI via our novel AI-instruments, rather than as views or controls with fixed, hand-designed and hard-coded options. AI-instruments can also be grounded in other instruments, enabling exploration of the space of related instruments (Figure 6 (3)).

5 Examples of AI-Instruments

To assess the viability of our AI instrumental model, study its differences with existing GUIs and tease out its value compare to existing chat-based AI interaction, we built a technology probe [34] with four different instruments, grounded in the literature: Fragments, Generative Containers, Transformative Lenses and Fillable Brushes. We describe below how this set of instruments surface the principles of our AI-instrumental model, as well as offer complementary interaction capabilities and affordances.

5.1 Fragments

Fragments build on the concept of *Attribute Cards* introduced in Object Oriented Drawing [74], as well as Side View's notions of *Variations* and *Parameter Spectrums* [67, 68], by using a large language model to extract multiple conceptual dimensions that may be plausibly implied by a prompt.

Fragments reify an initial prompt used to generate text or image into a set of attribute cards, of the format **`[type, value]`** (where **`type`** is the category of the extracted dimension, and **`value`** is the extracted value within that dimension—such as `[tone, enchanting]`, `[content, castle]` or `[style, illustration]`). Revealing these conceptual dimensions enables an initial reflection-in-intent, revealing the latent structure of the prompt as seen by the AI model. Commercial software such as Adobe Firefly [1] offers a similar capability as tags, enabling users to select them from a side panel for subsequent image generation. Applying the principle of reification to tags and turning them into cards affords three core novel interactions as illustrated in Figure 7.

First, users can reveal fragments via a long press on the content. Fragments are fully reified as interactive instruments and are dynamically generated—hence open-ended and nondeterministic—in contrast to the fixed, hand-crafted, and hard-coded controls supported by prior work (e.g. [35, 67, 74, 75]). Second, via drag and drop, users may remove fragments (by dragging them away), or add new fragments onto existing content in the work space. Adding or removing fragments triggers regeneration of the content. Third, to further support reflection, fragments offer suggestions on demand. By tapping on ···, users can generate new variations from any fragment; these suggestions appear in a column below the specific fragment. Users can also invoke additional suggestions for more types of fragments, which are then appended to the row of fragments.

These three core mechanisms support a workflow where, as users work with multiple images in their workspace, they can explore the effect of different fragments via drag-and-drop to ground one image generation into an aspect of another.

Fragments use the affordance of attribute cards to break down and reify a complex intent into manageable pieces, each having distinct type and value, that enable people to work with these as more-or-less independent and composable, "pieces of intent." This also encourages a workflow where users can surface useful fragmentary concepts surface that become reusable and specialized instruments in their own right. Such fragments are then available for reapplication to other pieces of content, or even reuse in a different context.

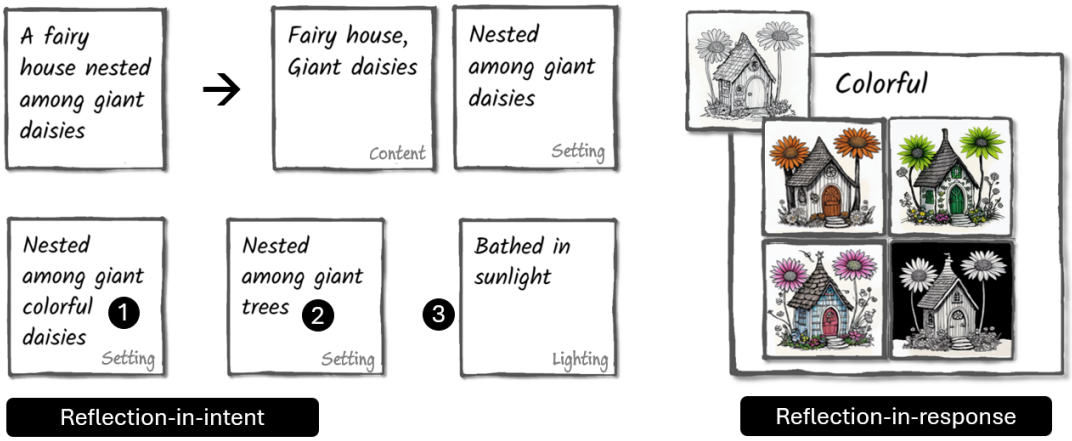


Figure 5: Reflection-in-intent enables users to gain awareness of the possible formulations of their intent while reflection-in-response enables users to assess the space of possibilities of the outputs generated by the model given an input. These aspects may help users address the challenges of intent disambiguation, resolution and steering.

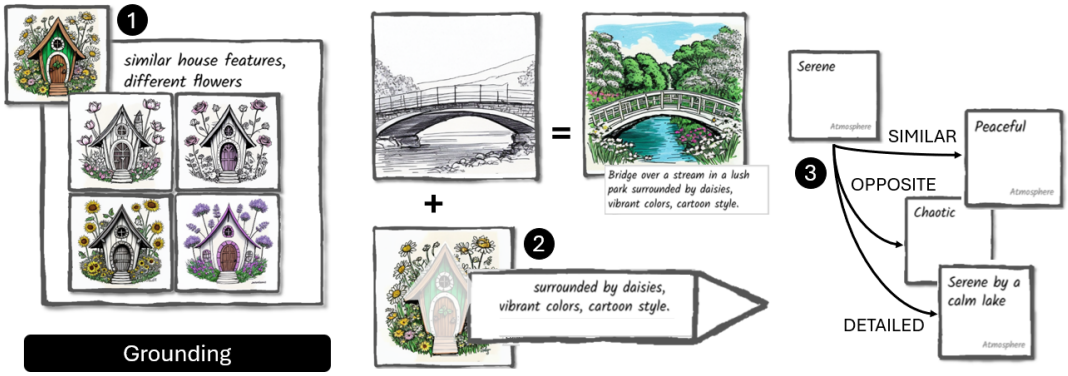


Figure 6: Grounding an instrument such as a generative container with an example enables to refer to features to preserve or alter in simple worlds by leveraging AI segmentation (1). Grounding an instrument such as a fillable brush in a specific aspect of an example, for example by selecting a region and extracting its style (2), enables users to use and apply it to other inputs without the need to articulating it in words. The principle of grounding also applies to instruments themselves such as deriving fragments from an example one (3).

While in principle we could have pursued a design that generated many fragments as automatic suggestions associated with each piece of content, such an approach would introduce clutter and risk overwhelming the user with the "decision paralysis" of too many choices. Our design therefore surfaces only a few fragments at a time, and only in a post-hoc manner upon explicit invocation by the user. Further, we present these newly-invoked fragments in an organized fashion, with two orthogonal dimensions of exploration on demand, by keeping dimension type in horizontal rows of cards, and value variations in vertical columns beneath these.

5.2 Transformative Lenses

Transformative Lenses re-envision the Toolglass and Magic Lens interaction technique [8] as a layered instrument that can be coupled with a generative prompt.

Layering a Transformative Lens on top of content uses such a prompt to generate a new image that synthesizes the lens and the content. Likewise, a specific piece of image content can be used on top of a lens to recombine the two. Such layerings can be positioned and manipulated to chain multiple effects together. As illustrated in Figure 8, users can leverage lenses to take a piece of content (e.g. a sketch of a suspension bridge), and then re-compose this content within a wider backdrop scene (a city skyline), or even apply a new specific style to the results with a single interaction (e.g. a heavy, black-lined graphic novel style).

More generally, depending on how the user layers Transformative Lenses and image content, lenses can support image completion from a small piece of content, synthesis and composition of multiple pieces of content into a new image, or regeneration of the underlying image. Note also that blank lenses (which have no image content, but do contain a prompt) can be used. For example, a

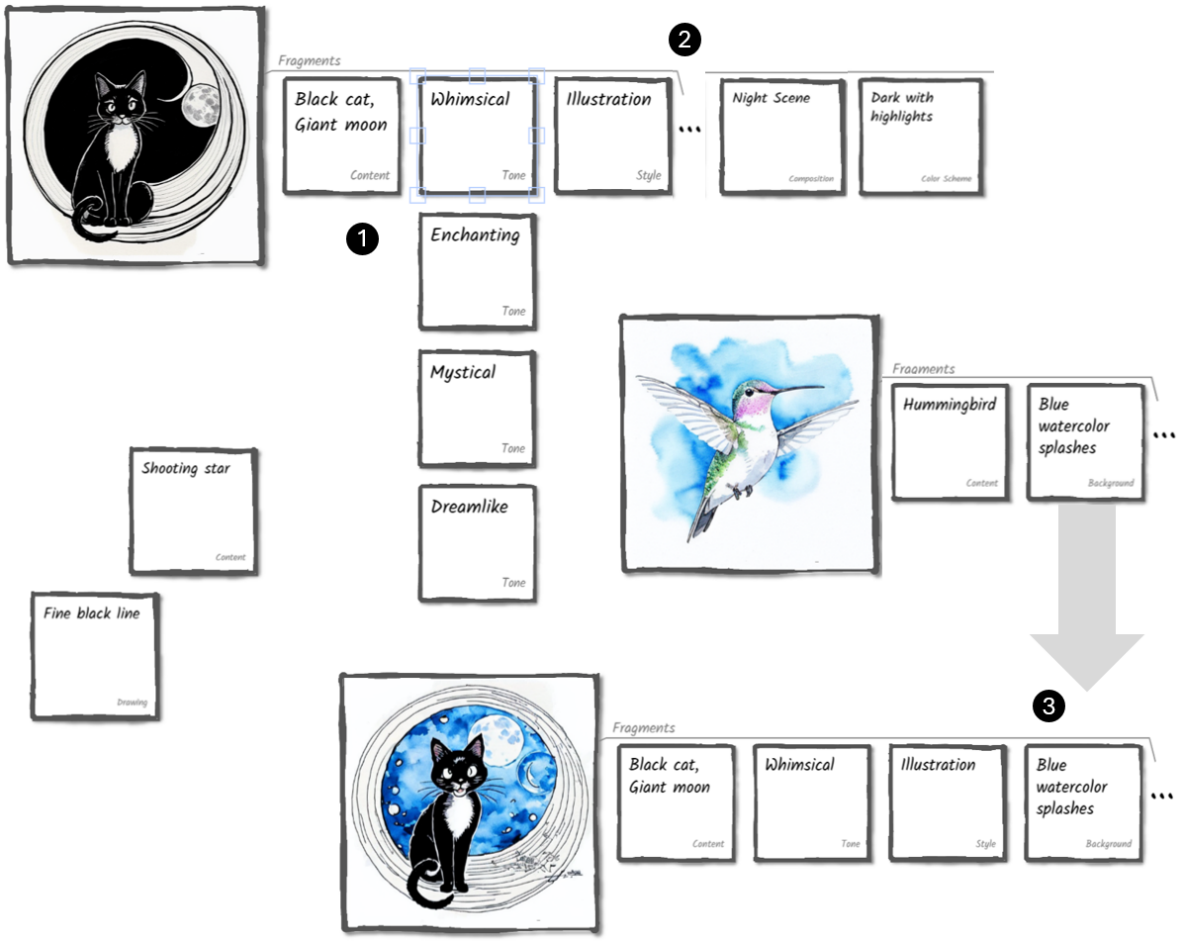


Figure 7: Users can expand Fragments with variations of parameter values (1) in vertical columns, or request more suggestions for dimension types (2) at the end of the row. Users can further reuse and transfer Fragments to other content via drag-and-drop(3).

blank-lens backdrop generated afforded outpainting-like operation —but here steered by the lens's prompt—to "complete" a scene from an existing piece of image content.

Users can freely drag, reposition, and resize both content images and lenses, layering them over each other to chain transformations, reflect on the results, and experiment with different combinations. This property also may encourage users to break down their intent into multiple lenses, which can then be applied to multiple pieces of content (grounding). Note that image recomposition and dynamic regeneration occurs after a 2-second idle time to avoid triggering constant image regenerations during dragging or resizing operations. As users may wish to adjust content under a lens post-generation, the lens is temporarily faded out in the background when the mouse pointer enters it.

Complementary to the Fragments described in the previous section, Transformative Lenses afford the design consideration of breaking down the output into pieces (whereas fragments focus on the prompt intent). People can control the composition of images by just moving and layering elements in relation to the lens, limiting the need for precise selection, and encouraging rapid iteration & experimentation with compositions. However, unlike an undo operation, removing (or otherwise reverting) the layering of Transformative Lens and image-content elements triggers re-generation, and will always lead to a slightly different rendering.

5.3 Generative Containers

Designers use moodboards [12], storyboards [29], and other techniques for presenting small-multiples in galleries [26, 49, 67] to illustrate and explore a space of possible creative directions. Structured generation of those alternatives [65] allows rapid exploration of design spaces, and techniques to highlight similarities and differences [24] facilitate the selection, refinement, and comparison of multiple responses.

As shown in Figure 9a, *Generative Containers* provide an AI-instrument that encapsulates these notions using a prompt—shown in the container's header—that is closely associated with a 2x2 small-multiple grid of generated image results. Users can then enter or

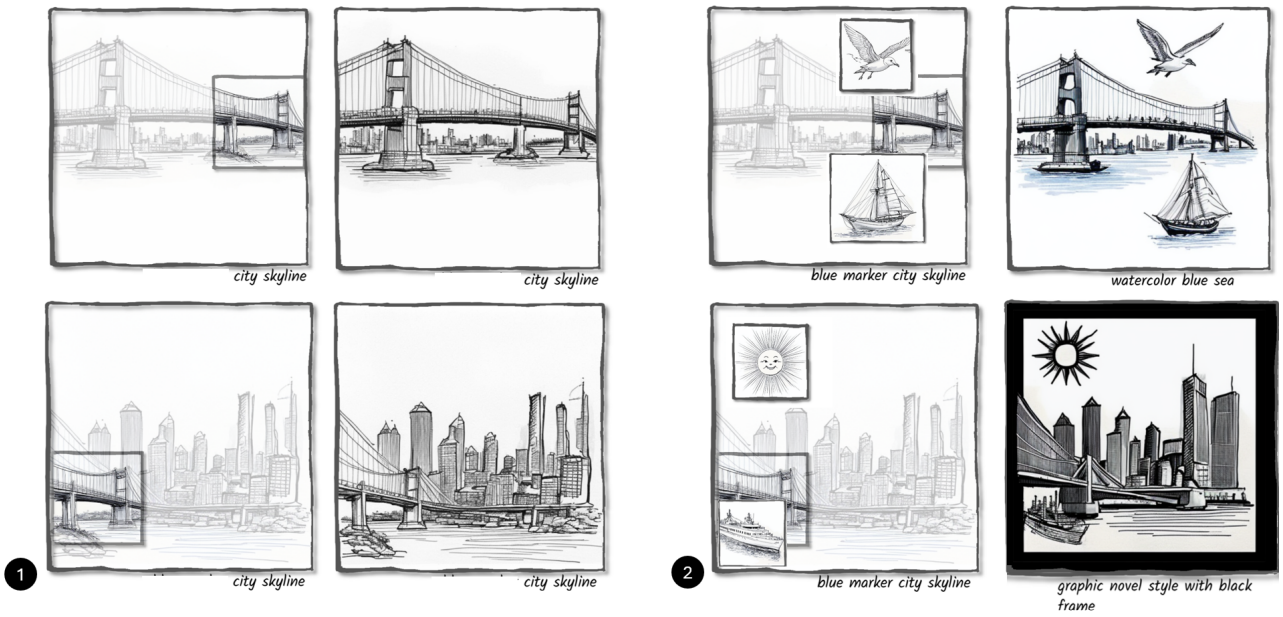


Figure 8: Transformative Lenses are placed over initial content, enabling users to "complete" illustrations from pieces of content (1). When users add elements to their composition, lenses regenerate to integrate it (2).

edit the prompt, or drag and drop example content—or even another instrument, such as a Fragment—onto the Container to ground it and generate a new small-multiple set of results.

We designed Generative Containers to enable reflection-in-response, allowing users to quickly get a visual sense of the range of responses a single prompt might produce. And by using Generative Containers to generate different variations of fragments, for example to obtain more concrete image editing suggestions from a high-level intent (Figure 10 left), generative containers also enable reflection-in-intent.

In our current implementation, the Generative Containers probe supports generation of four different variations (in a fixed 2x2 grid). Further, each container is presently limited to a single grounding example as input. However, users can create multiple containers and reuse results by dragging and dropping from one to another. In this way Containers afford adding details and varying the prompt to generate a range of example images, encouraging multiple cycles of iteration. Recombining and chaining these together effectively results in a longer, refined prompt that integrates the series of changes from prior interactions. Furthermore, one could expand the Generative Container instrument with other representations beyond our 2x2 grid, such as the dimension plots or stacked vertical dimension grids [65].

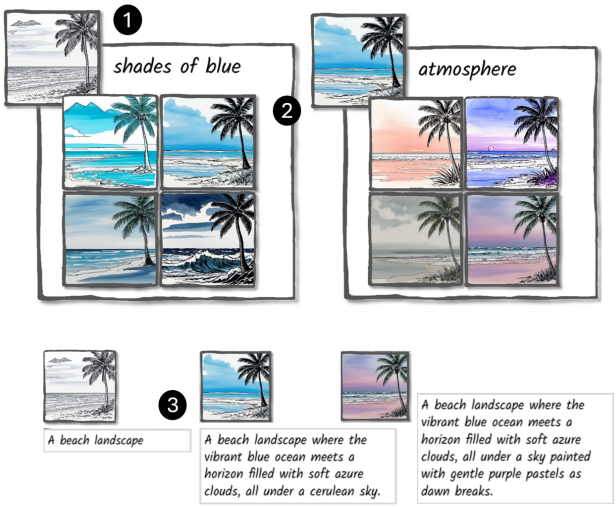
5.4 Fillable Brushes

Fillable Brushes, as illustrated in Figure 9b, offer an AI-instrument with the semantics of an "intelligent paint brush" for style transfer scoped to a particular spot on an existing image.

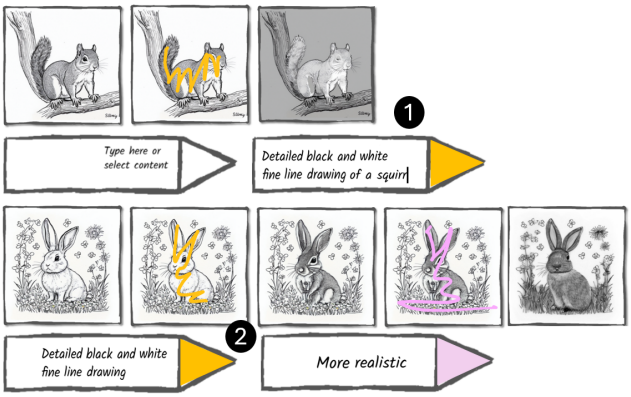
While previous work has explored brushes that can encapsulate and integrate deterministic modes and commands [58], our Fillable Brushes instrument applies encapsulated AI-prompts onto content in an intelligent manner as the user scrubs over it with their pen, finger, or other pointing device. And in contrast to the post-hoc notion of Fragments described above, Fillable Brushes apply a brush onto content, structured as an AI-Instrument "command" with a prefix (as opposed to postfix) syntax [13]. This offers a familiar interaction model from the way that a highlighting tool turns selected text yellow in a document editor, for example.

Encapsulating a prompt or image into a brush to define its function is a powerful interaction techniques to control scope of selection, as demonstrated by Runway motion brushes [60]. Applying additional principles of our model, enables users to also "fill" (ground) an empty brush by using existing content as an example, as if the instrument were a color picker that picks up key semantic attributes of the content rather than just its "color." The prompt encapsulated by the Fillable Brush is then automatically populated with descriptive words via generative AI, which the user can further edit if desired. This can be particularly helpful when users want to style something "like this" even when they may lack the vocabulary to describe its visual style. Our Fillable Brushes technology probe supports both content and/or style extraction. Turning a brush into a persistent object on screen, enables combining brushes together by drag and drop. Brushes can also be applied multiple times to the same content to emphasize a particular prompt in the result.

While Fillable Brushes enable users to specify the scope of intent with a high granularity, this does not necessarily require high precision: our implementation leverages the AI-powered Segment Anything Model (SAM) [40], which enables users to make approximate selections (i.e. rather than a precise and tedious lasso selection) to indicate an image element. The source content plus the approximate selection (as a set of reference points) is then converted into a precise object selection by the segmentation model.



(a) Generative containers enable users to explore possibilities on concrete or abstract dimensions (1). Containers also afford complex exploration paths by reusing the output of one container as the input of another one (2), resulting in refining intent (3).



(b) Brushes can extract aspects of content difficult for users to articulate in words, such as drawing style, making it reusable and editable (1). Combined with the selection afforded by brushes, this enables to apply aspects such as style to portions of images (2).

Figure 9: Generative Containers and Fillable Brushes support different types of content creation tasks. Containers promote the exploration of multiple ideas in parallel, while Brushes offer precise direct manipulation for steering generation. Providing users with both of these AI-instruments enables them to conduct many different activities involved in content creation, enabling interweaving of both divergent and convergent thinking activities.

5.5 Generated Instruments and Meta-Instruments

Beyond the concept of instruments, the instrumental interaction model [5] also refers to the concept of meta-instruments, in which *"instruments operate on instruments"*. As hinted at in earlier sections, using instruments on other instruments can be particularly useful to derive or compose instruments from the "task detritus" [41] already produced in the user's workflow and experimentation with other instruments. Using generative containers on Fragments, for example, can help users navigate the degree of abstraction, turning a vague idea into a set of concrete modifications (Figure 10 left).

However, such generation loops (instruments that generate content, generating instruments that generate other instruments generating content...) could potentially lead to an unwieldy number of elements in the interface. To organize but also generate collections of instruments, we devised a type of meta-instrument we call **Palettes**.

Akin to menus and containers available in GUIs today, Palettes enable storage and/or generation of different sets of instruments and content if desired. These afford abstraction and generalization of instrumental controls from collected pieces of content that can then serve as examples or generative seeds (Figure 10 right). Palettes of diverse instruments can balance the different affordances and properties of each instrument to provide rich content creation support. They can also help people get past the "cold-start" problem in complex creative design work, by beginning with examples, other pieces of existing content, or past work-artifacts to help overcome so-called "writer's block" or "blank canvas" effects of starting from nothing.

6 Implementation

Overview: Our system and all AI-Instruments technology probes were implemented on a web-based platform. We use Javascript and HTML with the fabric.js [38] library for the front-end, and a Node.js [22] server for the back-end managing content and files as well as coordinating communication with the generative AI models. For the user interface design, we chose to use a sketched user interface look and feel, to encourage our study participants to focus on the concepts rather than the surface details of their specific instantiation in the UI [12].

Leveraging Generative AI models: We use the OpenAI GPT-4o [53] model for text transformations and image analysis, and a local Stable Diffusion [63] server with a custom processing pipeline for image generation. The AI-Instruments use GPT-4o for analysis of provided input (e.g., for turning provided visual content into a text prompt, analyzing the contents of part of the workspace canvas). For image generation, we use multiple stacked ControlNet [80] models with Stable Diffusion to steer the generation of visual content. To preserve aspects of the source input, we use a combination of *Depth*, *Canny Edge*, and *Scribble* ControlNet models, while for preserving art/rendering styles (e.g., the sketch-based output) we use the *Reference* ControlNet model. Depending on the type of image generation, we vary the weight of each ControlNet model (e.g., increasing weight to emphasize content preservation, or decrease weight of another ControlNet to reduce affect of reference style transfer). We use image masks to selectively control which areas are changed or kept, apply inpainting/outpainting scripts, and adjust other parameters such as CFG scale, denoising strength, and control mode.

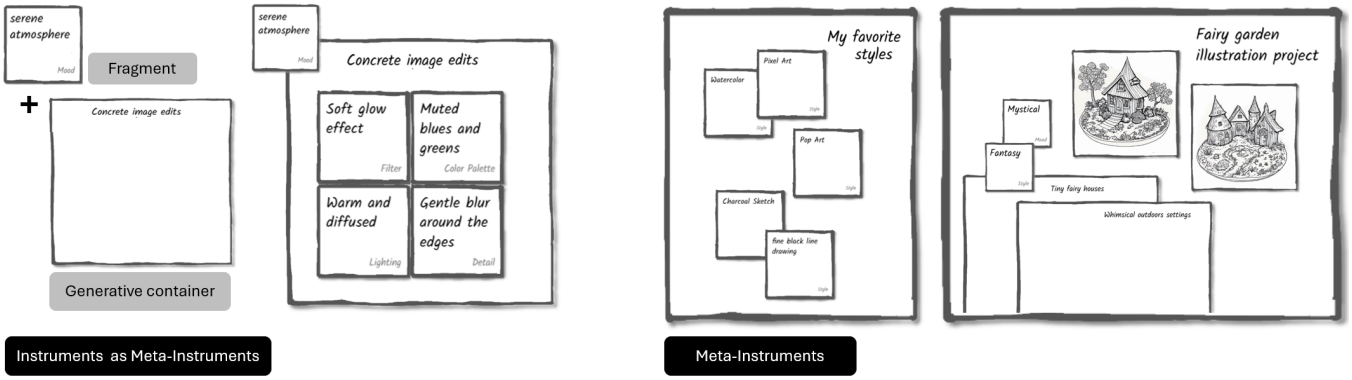


Figure 10: Instruments can be used as Meta-Instrument: operate on each other to create related instruments, for example for making a fragment more concrete (left). Specific Meta-Instruments such as palettes (right) can help user organize sets of instruments for easier retrieval and reuse, or, even help generating collection of instruments for a certain task.

Building AI-Instruments: We designed a pipeline that can orchestrate the access and requests to the different LLM and diffusion server instances to generate results. Key functionality is wrapped in modules, such as for encapsulating prompts to communicate with one or more models (by using model chaining) to perform a specific task. Each AI-instrument then uses a number of these modules for modifying the input or generating new content bases on the user's performed action with the instrument:

- ***Fragments*** instruments include modules for (1) prompt decomposition which takes a text or visual input and makes a GPT-4o request to generate fragments (returned as collections of `[type, value]` pairs), (2) fragment extension which takes a prompt and the existing fragments and requests additional fragment dimensions, (3) fragment variation which takes the fragment and parent prompt/content (if applicable) and generates variations of that fragment, and (4) prompt composition which takes a prompt, a modification to the fragments, and returns a modified prompt. The result from the prompt composition is then used to create an updated image with the Stable Diffusion + ControlNet pipeline.
- ***Transformative Lenses*** use a module for composition of the prompt (merging prompts from all source images covered by the lens), before then applying inpainting/outpainting, masks, and ControlNet models to generate the resulting image.
- ***Containers*** use a variation module, taking a prompt and a dimension, and requesting four variations along the provided dimension. Within the prompt we request visually diverse results. The resulting set of prompts is then sent to SD+ControlNet to generate the final set of four images in the container.
- ***Fillable Brushes*** are implemented to either emphasize style or content variations, depending on the intent of the user, which we support by varying the weight of the ControlNet models (e.g., higher reference ControlNet weight for changing the visual style, or increasing weight of Canny-edge/Depth ControlNet to preserve existing content). When the brush is applied, we perform a segmentation of the source content by feeding the stroke path as control points into Segment Anything [40], which results in a segmentation mask of the dominant object(s) selected with the brush stroke. We then use GPT-4o to craft a combined prompt given the source image(s), the segmented content, and the original prompt. Finally, we send this generated prompt together with the source image and segmentation mask to the Stable Diffusion server, using the ControlNet inpainting method.

7 Study

We conducted a qualitative user study with 12 participants to gather their insights on AI-instruments compared to traditional prompting. Participants completed image generation and editing tasks with our four technology probes as well as an initial chat-based prompting probe, to help them tease out pros and cons of these two different interaction models. We analyzed their comments to understand the perception of the principles of reification, reflection and grounding, as well as capture insights on the Human-AI interaction challenges (listed in Section 2.1) addressed by different instruments.

7.1 Procedure

Participants completed a 60-minute study in a quiet room on a computer running our technological probes and a study form. After obtaining informed consent, we collected basic demographic information, then requested participants to complete a set of tasks with our technological probes. All participants first completed tasks with a chat-based prompting probe using the same image generation model as the instruments in order for us to confirm their familiarity with prompting and to also provide them with a baseline for the image generation model with use. Before each instrument, participants watched a video demonstration explaining functionalities and modalities of interaction. After this video, participants used the technological probes to complete 2 to 3 tasks such as generating an image and changing its style. We provided example content for each task, but encouraged participants to generate their own content and try different ideas. The experimenter only interacted

with participants during this phase of the study to clarify functionalities and interaction if needed. After each set of tasks, participants reflected on one key advantage and on key inconvenient of this specific instrument compared with the chat-based prompting interfaces. Since we aimed at gathering qualitative insights on the overall interaction model behind instruments rather than compare instruments against each other, all participants completed the tasks in the same order. After tasks completion, we described five types of generic tasks, asking for each to select the best interaction technique. Participants also entered the rationale for their choice. At the end of the study, participants received a $50 gratuity for their time. The study protocol was reviewed and approved by the Microsoft Research ethics review board.

7.2 Participants

We recruited 12 participants (8 men, 3 women, 1 non-binary) via mailing lists in a large organization. We selected participants with weekly interaction with generative AI systems (ChatGPT, stable diffusion, etc) for creating content. As we aimed at gathering insights on overarching interaction principles for AI-instruments (across many domains), we opted for selecting participants with interesting in different types of content generation. Our participant pool included users interested in authoring short text snippets such as emails, long structured documents such as reports, structured text such as tables, programming code and web-pages, visual artifacts such as images, and multimodal artifacts with text image and charts such as presentations. Note that none of our participants generated audio or video.

7.3 Material and Analysis

We collected the salient advantage and salient weakness for each instrument compared to prompting. Participants experienced the following probes: 1) chat-based prompting, 2) fragments, 3) containers, 4) lenses, and 5) brushes. To encourage participants to think of different aspects of content generation, we asked them their preferred interaction technique (along with their rationale) for five different tasks.

T1 Combining content: merging pieces of content together
T2 Splitting content: extracting a piece of content
T3 Iterating on content: editing an aspect of content
T4 Editing by example: transferring content or style
T5 Expanding content: adding new material to existing content

We coded a total of 156 statements from our participants to gain insights on their perception of our model's principles and to assess how AI-instruments (un)successfully addressed the five challenges described in section 2.1. A portion of these comments (28/156) also revealed limitations of our technical probes (the codebook is available at https://hugoromat.github.io/ai_instruments/).

7.4 Insights on Model Principles

Reification of intent. All 12 participants reacted positively to the principle of reifying intent into AI-instruments. Participants valued that AI-instruments enabled them to shift the focus from the prompt to the outcome. P6 commented that *"I can just click on the fragments instead of typing it out and focus on the final output instead."*. P10 noted it was helping them with the iterative process: *"with prompting it makes me think of the prompt, but having [fragments] already in front of me can make it easier for me to make a decision of what i want."*. They also valued the **direct interaction** afforded by AI-instruments: *"[with fillable brushes] you can interact with the images that are generated directly, rather than [modifying] the image only from prompting."*

A few participants also outlined the value of reification for storing and reusing prompts. P9 commented on Fillable Brushes *"[...] I would be able to create a [brush] with the thing that I wanted to pull out and then apply/store it however I wanted."*, and P11 on Transformative Lenses *"I like the idea of creating and saving a lens and applying it consistently to different images for future uses"*.

All 12 participants also commented on **scope of selection** as a key advantage of AI-instruments over prompting (38 comments). Participants identified Fillable Brushes as enabling them to specify portion of an image while Transformative Lenses enabled them to combine multiple images together. With Brushes, participants emphasize the granularity of the selection. For example P1 noted *"I can highlight the part of the image that requires changing only, and it seems I can highlight at a very high granular level, such as a face of a dog."* P7 referred to this capacity for steering image generation as one can use masks in graphics editors *"I can control a more fine grained area/mask that I want to edit. It's so cool! "* P4 noted Lenses were particularly useful for adding elements iteratively: *"about how to incrementally add new items from an initial picture, the others [AI-instruments] are more for customize different picture styles"*. In addition, participants appreciated controlling image composition with Lenses, as P2 explained: *"Being able to specify the location of component objects is really helpful"*.

Reflection. All 12 participants pointed to **reflection-in-intent** as an advantage of AI-instruments in contrast to prompting (31 comments). This principle was particularly highlighted as a strength of Fragments (24/31 comments), as P5 explained *"I could tell what the different aspects of the prompts were being split [...] and guess what the AI interpreted as something other than what I had in mind."* Participants also praised the benefit of generating variations for each fragment such as P11: *"the system generates ideas for you which you can implement, allowing you to add variables which you might not have originally thought of."*

All 12 participants also outlined **reflection-in-response** as an advantage of AI-instruments in contrast to prompting (31 comments). Generative Containers were mostly (26/31) cited for this ability. P5 explained that *"Usually I ask AI to give me a different version/example [of text], but this would allow you to choose the one you like without needing to prompt it again."* P2 valued this ability for iterating: *"[generative containers are a] great abstraction for deciding what to iterate on given multiple possibilities. Makes it easier to visualize or use results from previous steps"*.

Grounding. All 12 participants identified grounding as an advantage of AI-instruments over prompting (45 comments). It was especially noted valuable for images, as P4 explains *"apply styles of different pictures into other ones, sometimes the styles are hard to illustrate in prompting, since they are more abstract"*.

A majority of participants referred to grounding as a key advantage of Fillable Brushes. P8 referred to grounding for dealing with multiple items: *"[...] having the flexibility to copy any kind of prompt*

on the brush is good if I have similar kind of images to replicate." P9 referred to grounding for iterating on a single item: *"[brushes] can give us a style for A (which we can store and ensure that it gets applied to all our later additions.)"*.

Model limitations. We gathered 20/156 comments pertaining to the model limitations.

Four participants described **limitations of GUIs** compared to chat-based conversational interactions with AI. For example, P11 described accessibility issues of relying on 2D interactions as opposed to the ability to rely solely on speech as is easily feasible with chat-based prompting interaction. P2 mentioned AI-Instruments would suffer from the same issues they encountered with GUIs: *"GUIs in general are more prone to bugs or unexpected behaviors than text interactions, which could lead to increased computation that the user doesn't actually want"*. And P10 noted that they felt they needed to perform many interactions *"compared to prompting, where I can just type something"*, P12 concluding that *"It's faster to get exactly what I want with prompting"*. These comments do not address fundamental limitations of the instrumental model per say but rather highlight that users may sometimes prefer a single albeit limited modality of interaction. These insights hints at the need to integrate chat-based prompting interaction more tightly with instruments. A straightforward avenue for this, is to provide on-demand access to underlying textual prompts generated by instruments.

P6 noted that *"It might be better to write a new prompt when I need to make major modifications to the output"* as a drawback of Lenses. This echoes the sentiment of three other participants in that the grounding capabilities of AI-instruments such as Lenses and Brushes enable effective content generation steering but may hinder divergent content generation. However, this is also balanced by the strengths of other AI-instruments such as Generative Containers that many participants praised for *"creative and exploratory settings."* (P12). These insights hints at the need to **provide multiple instruments** to users in content generation tools.

We devised our four AI-instruments to cover different facets and tasks involved in content creation. For example, Containers supports iterative exploration by reusing pieces of content in other Containers, while Fragments enables it by selecting different aspects to vary. The different design decisions tied to the affordances of each instrument led to different task support, especially with regard to navigating creation history. This caused occasional frustration for Fragments or Lenses *"Old edits get lost when new edited are prompted"* (P7). Two participants reflected on this strength of chat-based prompting interfaces to inherently provide a trace of all the prompts and results made in a temporal manner. Especially in the context of non-deterministic outputs, where the next piece of content may be less satisfactory than the prior one, these insights suggest that more thoughts need to be devoted to **systematically surface history** within each instrument.

Note that a majority of aspects participants did not value for AI-instruments (28/48) pertained to the limitations of our technology probes rather than limitations of the model. Many of these comments referred to usability issues such as the fact that Lenses regenerated too early (or too late) or exclusively had a square aspect ratio, as well as visual look and feel of the probes *"color the pens maybe? easier to organize and distinguish them"*(P7).

7.5 Challenges Addressed by AI-instruments

C1 Intent formulation. Participants highlighted **grounding** to address intent formulation on images, especially as instantiated in Fillable Brushes for extracting and transferring styles. P7 summarized how Brushes help formulate intent with both scope of selection and direct manipulation: *"I can control a more fine grand area/mask that I want to edit. It's so cool! I can also copy the style/content from a different image and directly apply it to a different image without having to find the accurate words to describe them."*

C2 Intent disambiguation. Participants identified **grounding and reflection** as core principles helping them disambiguate their intent. They explained that the reflection-in-intent surfaced in Fragments was helping them identify context and details to refine their intent. P5 also noted that Fragments enabled to experiment and gain an understanding of how the model worked *"allow[ing] me to see how the prompts were being isolated and used"*. A few participants also explained that Brushes helped with intent disambiguation by capturing one by example and expressing it words. P1 summarized: *"one advantage [of brushes] is that it can identify a style, even though I cannot articulate the style well, this is especially helpful to circumvent prompt engineering."*

C3 Intent resolution. Participants outlined the principles of **reification and reflection** as most useful for resolving intent. Participants explained that reification enabled them to explore different paths by iterating on different images: *"this lets you build off of previous iterations and you can create different styles for the same image until you are happy with one"*. They highlighted the benefit of reflection-in-response of Generative Containers to help them explore possibilities:*"I don't have to think too much about what I want, which is helpful"* (P10), *"Very easy to explore creative options, it helps to get a better idea of what you like"* (P12). P7 valued the ability to start from high-level prompts and follow up by making a selection *"It it easy to generate images with a vague description, and offer options [to explore]"*. A few participants also mentioned the benefit of reflection-in-intent of Fragments to experiment with options: *"[Fragment] also suggests some dimensions to change the picture which I might have not thought about. Like the elephant in this case."* (P9).

C4 Steering. All participants mentioned the benefit of instruments over chat-based prompting for steering content. Participants noted that direct manipulation and scope of selection afforded by **reification** were critical in editing portions of content, in conjunction with **grounding** to capture and transfer aspects of one content to another. Many participants outlined the value of Brushes and Lenses to generate masks for steering content generation *"masking [with lenses] allows me to personalize smaller parts of the image compared to generating a completely new one"* (P10), *"this [brush] is an advanced way to mask out single things within an image to regenerate"* (P3). P6 summarized the benefits of these instruments compare to chat-based prompting: *"[with brushes] I can quickly make modifications to the image instead of writing a prompt from scratch for every modification. I can focus on the subject I'm interested in."*

C5 Workflows. Participants perceived that AI-instruments supported two workflows that were currently hard to achieve with chat-based prompting interaction: **non-linear exploration** and **iterative content generation** by chaining prompts together (e.g. steering). For exploration, participants each favored different instruments. P6 favored Fragments *"I feel this [fragments] makes it easy to try out different examples that I might be interested in. I can just click on the fragments instead of typing it out and focus on the final output instead."*. P3 described Containers allowing *"more freedom to explore more options. Downside to prompting is that you need to know what you want."* P8 valued *"trying different lenses on the same component creating a scene is easy way to experiment and merge different style in one"*. P8 also referred to the grounding capabilities of Brushes as useful to experiment with the collection of examples offered *"different kinds of image styles to choose from for copying styles or format"*. For iterative content generation, participants refer to the same advantages AI-instruments offer than for steering.

8 Discussion and Future Work

We first discuss how our model revisits and extends the classic instrumental interaction model, then discuss the application of AI-instruments to different forms of content, outlining future work.

8.1 Revisiting and Extending the Instrumental Interaction Model

The instrumental interaction model [5] is based on three core principles [7]:

- *Reification of commands* refers to the principle of turning systems functionalities into interactive graphical objects in the interface,
- *Polymorphism* refers to the principle of applying commands to different types of objects enabling the interface designers to keep the number of interface objects relatively small,
- *Reuse* refers to the ability for users to reapply one command to different objects or apply different commands to one object, with the goal of limiting repetitive user input and/or navigation.

In this paper, we revisited and proposed to extend this model in the following ways:

(1) We extended the principle of reification from encompassing a limited set of commands defined for an application to include any intent user may express in natural language. We also unpack two key considerations of reification that one should consider when designing AI-Instruments: the scope of the instrument, and degree of abstraction. We propose to leverage the affordances of existing direct manipulation techniques to convey to users how to specify scope (e.g. select a portion with a brush vs resize the lens). To support users navigating different levels of abstraction of their intent, we propose to use AI-instruments themselves.

(2) We re-framed the concept of polymorphism in instrumental interaction, recasting it as *reflection*. This shifts instrumental interaction from a classic "direct manipulation" technique for graphical user interfaces to a modern AI-augmented technique. Reflection leverages the general concept-translation capabilities of LLM's to support an expansive notion of "polymorphism" without requiring the interface design and system architect to hand-code the parameters, controls, and nuances of how these are interpreted across a wide range of content types.

Further, by considering reflection from both the user (intent) and system (response) perspectives, we provide users with mechanism to explore both the design space of their intent (i.e. different formulations and disambiguation of intent), as well as the design space of the model response (i.e. different interpretations of user intent by the model). While such notions, in one sense, have been latent in interactive instruments all along, with generative AI many possible forms and interpretations of polymorphism—potentially even for niche or specialized workflows, formats, and types of content, if they are sufficiently represented in the training data for the model—can be made available for user reflection in AI-assisted content creation.

(3) The third principle of reuse is closely related to our principle of grounding. Grounding extends the notion of *reusing commands* to the capability of *extracting and reusing intent*—whether in terms of one aspect of user intent, a collection of multiple user intents, or other properties of content. This principle of grounding also encapsulates the ability to generate instruments from other instruments, characterized as *meta-instruments* in the nomenclature of instrumental interaction [5, 7].

(4) A further new challenge raised by AI-instruments is the need to balance the possibility of over-generating instruments, with the power to encapsulate many capabilities—at a high level of abstraction—within a single instrument. In contrast to classic handcrafted instrumental interaction, AI-generated instruments could potentially lead to an unwieldy number of objects in the workspace, if generation were left unchecked. However, we counterbalance this with strategies to compose instruments and organize them into collections (meta-instruments). Further, we can leverage the generative nature of AI to iteratively refine both content and (meta-)instruments—altering, summarizing, or abstracting instruments and content with each step—as another strategy to harness AI to express aggregated concepts at a high semantic level.

The insights we gained by building a set of technology probes and gathering initial perceptions of 12 content creators, suggest that the principles described in our model can be used inform the design of novel interaction techniques as well as assess existing ones. As we built each probe, it became clear that design decisions at lower-level, for example pertaining to the specific choice of interactions (e.g. click vs double-click) or their timing (e.g. idle time threshold triggering image generation), can lead to different experiences, especially when multiple probes are used in conjunction. While our model suggests overarching principles for AI-instruments, specific interaction is bound to differ as sets of these techniques are integrated into specific applications and adapted to different modalities and contexts [3, 46].

Additional considerations for integrating AI-instruments in applications include the expectations of users with regard to direct manipulation and instrumental interfaces, as well as when working with AI. For example, a fundamental principle of direct manipulation is ease of reversibility of user actions (e.g. if a fragment is removed from an image triggering a new generation, then added back; the image should revert to its prior state). In contrast, AI models are non-deterministic by nature (e.g. same input, different

output). While it is technically feasible to couple AI generation with history and versioning mechanisms to ensure the reversibility of operations, users' attitude on working with AI over the longer term, as well as specific use cases may impact this design decision.

8.2 Beyond Images, Applying AI-Instruments to Other Forms of Content

While extrapolations from participants must be taken with caution, four participants in our study related the use of AI-instrument to content they worked with every day. Interestingly two participants (P2 and P9) commented they would not see the use of instruments for tasks such as writing code. P9 summarized it as *"it [fragments] is a bit harder to use than prompting for tasks like maybe writing code. I would prefer fragmenting for images or plots"*. However, two different participants (P11 and P12) envisioned using AI-instruments for data analysis and writing (P11) and in the case of P12 leveraging grounding for operating at the artifact level: *"it would be so cool to auto-pick a style (words, design, etc) without me first having to decipher it, and then have it automatically apply to other content (writing, slides, etc)"*.

We experimented with a few of our technology probes (Fragments, Containers and Palettes) to work with textual content and found that our principles generally held. However, further exploration with different types of modality is likely to reveal additional design considerations for the instruments we proposed. Notably, a key issue to address for textual content as opposed to images is the effort required to consume a number of potential outputs (reflection-in-response). Integrating support for helping user skim and get the gist of similarities and differences between textual outputs in Generative Containers (by bolding portions of text or providing summaries or excerpts) would certainly be necessary when using instruments to work with textual documents.

Another aspect to address is the use of instruments for artifacts composed of multiple pieces of content (e.g. a slide composed of a title and image). Again, while we believe core principles hold for devising instruments to work with content at the artifact level, additional research is needed to delve into how to integrate different aspects of an artifact. For example, one could envision displaying Fragments from different scope of selection, enabling Fragments to operate at the entire slide level or on a subset such as title.

In the future, we plan to pursue these two research directions (designing AI-instruments for heterogeneous content and artifacts composed of multiple pieces of content), further assessing the generalizability of our interaction model and broadening the set of design considerations for AI-instruments.

9 Conclusion

We operationalized the theory of instrumental interaction for generative AI, with an in-depth unpacking of the principles of reification of user intent, reflection, and grounding. We argue that leveraging this re-appropriated and refined theory can drive the creation of a *new generation of expressive AI-Instruments* that afford better expression of intent, make it easier to discover what is possible, and provide powerful degrees of freedom for steering the generation towards the best possible results. Those new tools and instruments can truly leverage the polymorphic and non-deterministic behavior of generative AI models, unleashing new and empowering forms of expressive HCI+AI experiences.

Beyond our focus on AI-Instruments, theories play an important role in the advancement of our wider research field [27, 57]. Rogers argues that there is a need for theories as lenses bringing critical design characteristics into focus, and which can function as a generative source: providing "*design dimensions and constructs to inform the design and selection of interactive representations*" [56]. We hope that our work on operationalizing the theory of instrumental interaction for AI can inspire other new – and re-appropriated – theories to advance HCI+AI.

Acknowledgments

We thank the participants of our AI-Instruments user study for taking the time to provide feedback on our techniques, and the reviewers of this submission for their constructive suggestions to improve this research.

References

[1] Adobe. 2024. Firefly. https://www.adobe.com/products/firefly
[2] Shm Garanganao Almeda, J.D. Zamfirescu-Pereira, Kyu Won Kim, Pradeep Mani Rathnam, and Bjoern Hartmann. 2024. Prompting for Discovery: Flexible Sense-Making for AI Art-Making with Dreamsheets. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3613904.3642858
[3] Caroline Appert, Michel Beaudouin-Lafon, and Wendy E Mackay. 2005. Context matters: Evaluating interaction techniques with the CIS model. In *People and Computers XVIII—Design for Life: Proceedings of HCI 2004*. Springer, Springer, New York, NY, USA, 279–295.
[4] Yavar Bathaee. 2017. The artificial intelligence black box and the failure of intent and causation. *Harv. JL & Tech.* 31 (2017), 889.
[5] Michel Beaudouin-Lafon. 2000. Instrumental interaction: an interaction model for designing post-WIMP user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 446–453.
[6] Michel Beaudouin-Lafon, Susanne Bødker, and Wendy E Mackay. 2021. Generative theories of interaction. *ACM Transactions on Computer-Human Interaction (TOCHI)* 28, 6 (2021), 1–54.
[7] Michel Beaudouin-Lafon and Wendy E Mackay. 2000. Reification, polymorphism and reuse: three principles for designing visual interfaces. In *Proceedings of the working conference on Advanced visual interfaces*. Association for Computing Machinery, New York, NY, USA, 102–109.
[8] Eric A. Bier, Maureen C. Stone, Ken Pier, Ken Fishkin, Thomas Baudel, Matt Conway, William Buxton, and Tony DeRose. 1994. Toolglass and magic lenses: the see-through interface. In *Conference Companion on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '94)*. Association for Computing Machinery, New York, NY, USA, 445–446. https://doi.org/10.1145/259963.260447
[9] Aras Bozkurt. 2024. Tell Me Your Prompts and I Will Make Them True: The Alchemy of Prompt Engineering and Generative AI. *Open Praxis* 16, 2 (April 2024), 111–118. https://doi.org/10.55982/openpraxis.16.2.661
[10] 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 (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3586183.3606725
[11] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).
[12] Bill Buxton. 2007. *Sketching User Experiences: Getting the Design Right and the Right Design*. Morgan Kaufmann, Burlington. https://doi.org/10.1016/B978-0-12-374037-3.X5043-3
[13] William Buxton. 1983. Lexical and pragmatic considerations of input structures. *SIGGRAPH Comput. Graph.* 17, 1 (jan 1983), 31–37. https://doi.org/10.1145/988584.988586
[14] William Buxton. 1995. Chunking and phrasing and the design of human-computer dialogues. In *Readings in human–computer interaction*. Elsevier, 494–499.

[15] Tracy Diane Cassidy. 2008. Mood boards: Current practice in learning and teaching strategies and students' understanding of the process. *International journal of fashion design* 1, 1 (2008), 43–54.

[16] DaEun Choi, Sumin Hong, Jeongeon Park, John Joon Young Chung, and Juho Kim. 2024. CreativeConnect: Supporting Reference Recombination for Graphic Design Ideation with Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–25. https://doi.org/10.1145/3613904.3642794

[17] John Joon Young Chung and Eytan Adar. 2023. PromptPaint: Steering Text-to-Image Generation Through Paint Medium-like Interactions. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3586183.3606777

[18] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching Stories with Generative Pretrained Language Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3491102.3501819

[19] Gregory Thomas Croisdale, John Joon Young Chung, Emily Huang, Gage Birchmeier, Xu Wang, and Anhong Guo. 2023. DeckFlow: A Card Game Interface for Exploring Generative Model Flows. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 1–3. https://doi.org/10.1145/3586182.3615821

[20] Nicholas Davis, Chih-Pin Hsiao, Yanna Popova, and Brian Magerko. 2015. An enactive model of creativity for computational collaboration and co-creation. *Creativity in the digital age* (2015), 109–133.

[21] Umer Farooq, John M Carroll, and Craig H Ganoe. 2005. Supporting creativity in distributed scientific communities. In *Proceedings of the 2005 ACM International Conference on Supporting Group Work*. Association for Computing Machinery, New York, NY, USA, 217–226.

[22] OpenJS Foundation. 2024. Node.js. https://nodejs.org/en

[23] Charles Freeman, Sara Marcketti, and Elena Karpova. 2017. Creativity of images: using digital consensual assessment to evaluate mood boards. *Fashion and Textiles* 4 (2017), 1–15.

[24] Katy Ilonka Gero, Chelse Swoopes, Ziwei Gu, Jonathan K. Kummerfeld, and Elena L. Glassman. 2024. Supporting Sensemaking of Large Language Model Outputs at Scale. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 838, 21 pages. https://doi.org/10.1145/3613904.3642139

[25] James J Gibson. 1977. The theory of affordances. *Hilldale, USA* 1, 2 (1977), 67–82.

[26] Joshua Hailpern, Erik Hinterbichler, Caryn Leppert, Damon Cook, and Brian P. Bailey. 2007. TEAM STORM: demonstrating an interaction model for working with multiple ideas during creative group work. In *Proceedings of the 6th ACM SIGCHI Conference on Creativity & Cognition* (Washington, DC, USA) *(C&C '07)*. Association for Computing Machinery, New York, NY, USA, 193–202. https://doi.org/10.1145/1254960.1254987

[27] Christine A. Halverson. 2002. Activity Theory and Distributed Cognition: Or What Does CSCW Need to DO with Theories? *Computer Supported Cooperative Work (CSCW)* 11, 1 (March 2002), 243–267. https://doi.org/10.1023/A:1015298005381

[28] Björn Hartmann, Scott R. Klemmer, Michael Bernstein, Leith Abdulla, Brandon Burr, Avi Robinson-Mosher, and Jennifer Gee. 2006. Reflective physical prototyping through integrated design, test, and analysis. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology* (Montreux, Switzerland) *(UIST '06)*. Association for Computing Machinery, New York, NY, USA, 299–308. https://doi.org/10.1145/1166253.1166300

[29] 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. https://doi.org/10.1145/2858036.2858079

[30] Robert R Hoffman, Gary Klein, and Shane T Mueller. 2018. Explaining explanation for "explainable AI". In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 62. SAGE Publications Sage CA: Los Angeles, CA, SAGE Publications, Los Angeles, CA, USA, 197–201.

[31] Kasper Hornbæk and Antti Oulasvirta. 2017. What is interaction?. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 5040–5052.

[32] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 159–166.

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

[34] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 17–24. https://doi.org/10.1145/642611.642616

[35] Takeo Igarashi and John F. Hughes. 2007. A suggestive interface for 3D drawing. In *ACM SIGGRAPH 2007 Courses* (San Diego, California) *(SIGGRAPH '07)*. Association for Computing Machinery, New York, NY, USA, 20–es. https://doi.org/10.1145/1281500.1281531

[36] Robert JK Jacob, Audrey Girouard, Leanne M Hirshfield, Michael S Horn, Orit Shaer, Erin Treacy Solovey, and Jamie Zigelbaum. 2008. Reality-based interaction: a framework for post-WIMP interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 201–210.

[37] Peiling Jiang, Jude Rayan, Steven P. Dow, and Haijun Xia. 2023. Graphologue: Exploring Large Language Model Responses with Interactive Diagrams. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3586183.3606737

[38] Juriy Zaytsev, Stefan Kienzle, and Andrea Bogazzi. 2024. Fabric.js. https://github.com/fabricjs/fabric.js

[39] Tae Soo Kim, Yoonjoo Lee, Minsuk Chang, and Juho Kim. 2023. Cells, generators, and lenses: Design framework for object-oriented interaction with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 1–18.

[40] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. https://arxiv.org/abs/2304.02643v1

[41] David Kirsh. 1995. The intelligent use of space. *Artif. Intell.* 73, 1–2 (feb 1995), 31–68. https://doi.org/10.1016/0004-3702(94)00017-U

[42] Amy J Ko, Brad A Myers, and Htet Htet Aung. 2004. Six learning barriers in end-user programming systems. In *2004 IEEE Symposium on Visual Languages-Human Centric Computing*. IEEE, IEEE, New York, NY, USA, 199–206.

[43] Jingyi Li, Eric Rawn, Jacob Ritchie, Jasper Tran O'Leary, and Sean Follmer. 2023. Beyond the Artifact: Power as a Lens for Creativity Support Tools. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 1–15.

[44] Vivian Liu. 2023. Beyond Text-to-Image: Multimodal Prompts to Explore Generative AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3544549.3577043

[45] Vivian Liu and Lydia B Chilton. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–23. https://doi.org/10.1145/3491102.3501825

[46] Wendy E Mackay. 2002. Which interaction technique works when? Floating palettes, marking menus and toolglasses support different task strategies. In *Proceedings of the Working Conference on Advanced Visual Interfaces*. Association for Computing Machinery, New York, NY, USA, 203–208.

[47] Wendy E Mackay and Anne-Laure Fayard. 1997. HCI, natural science and design: a framework for triangulation across disciplines. In *Proceedings of the 2nd conference on Designing interactive systems: processes, practices, methods, and techniques*. Association for Computing Machinery, New York, NY, USA, 223–234.

[48] Atefeh Mahdavi Goloujeh, Anne Sullivan, and Brian Magerko. 2024. Is It AI or Is It Me? Understanding Users' Prompt Journey with Text-to-Image Generative AI Tools. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3613904.3642861

[49] J. Marks, B. Andalman, P. A. Beardsley, W. Freeman, S. Gibson, J. Hodgins, T. Kang, B. Mirtich, H. Pfister, W. Ruml, K. Ryall, J. Seims, and S. Shieber. 1997. Design galleries: a general approach to setting parameters for computer graphics and animation. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '97)*. ACM Press/Addison-Wesley Publishing Co., USA, 389–400. https://doi.org/10.1145/258734.258887

[50] Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. 2024. DirectGPT: A Direct Manipulation Interface to Interact with Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3613904.3642462

[51] Donald A Norman. 1986. Cognitive engineering. *User centered system design* 31, 61 (1986), 2.

[52] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13.

https://doi.org/10.1145/3173574.3174223
[53] OpenAI. 2024. GPT-4o System Card. https://openai.com/index/gpt-4o-system-card/
[54] Xiaohan Peng, Janin Koch, and Wendy E. Mackay. 2024. DesignPrompt: Using Multimodal Interaction for Design Exploration with Generative AI. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 804–818. https://doi.org/10.1145/3643834.3661588
[55] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988* (2022).
[56] Yvonne Rogers. 2004. New Theoretical Approaches for Human-Computer Interaction. *Annual Review of Information Science and Technology (ARIST)* 38 (2004), 87–143. ERIC Number: EJ678114.
[57] Yvonne Rogers. 2012. *HCI Theory: Classical, Modern, and Contemporary* (1 ed.). Morgan & Claypool Publishers.
[58] Hugo Romat, Nicolai Marquardt, Ken Hinckley, and Nathalie Henry Riche. 2022. Style Blink: exploring digital inking of structured information via handcrafted styling as a first-class object. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14.
[59] Karl Toby Rosenberg, Rubaiat Habib Kazi, Li-Yi Wei, Haijun Xia, and Ken Perlin. 2024. DrawTalking: Towards Building Interactive Worlds by Sketching and Speaking. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–8. https://doi.org/10.1145/3613905.3651089
[60] Runway. 2024. Motion Brushes. https://youtu.be/zQ3fQt8swEI
[61] D. A. Schön. 1992. Designing as reflective conversation with the materials of a design situation. *Know.-Based Syst.* 5, 1 (mar 1992), 3–14. https://doi.org/10.1016/0950-7051(92)90020-G
[62] Ben Shneiderman. 2007. Creativity support tools: accelerating discovery and innovation. *Commun. ACM* 50, 12 (Dec. 2007), 20–32. https://doi.org/10.1145/1323688.1323689
[63] StabilityAI. 2024. CompVis/stable-diffusion · Hugging Face. https://huggingface.co/CompVis/stable-diffusion
[64] 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 (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3613904.3642754
[65] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–26. https://doi.org/10.1145/3613904.3642400
[66] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM, San Francisco CA USA, 1–18. https://doi.org/10.1145/3586183.3606756
[67] Michael Terry and Elizabeth D. Mynatt. 2002. Recognizing creative needs in user interface design. In *Proceedings of the 4th Conference on Creativity & Cognition* (Loughborough, UK) *(C&C '02)*. Association for Computing Machinery, New York, NY, USA, 38–44. https://doi.org/10.1145/581710.581718
[68] Michael Terry and Elizabeth D. Mynatt. 2002. Side views: persistent, on-demand previews for open-ended tasks. In *Proceedings of the 15th Annual ACM Symposium on User Interface Software and Technology* (Paris, France) *(UIST '02)*. Association for Computing Machinery, New York, NY, USA, 71–80. https://doi.org/10.1145/571985.571996
[69] 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*. 1243–1252.
[70] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. 2022. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*.
[71] Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. 2023. Concept Decomposition for Visual Exploration and Inspiration. *ACM Trans. Graph.* 42, 6, Article 241 (Dec. 2023), 13 pages. https://doi.org/10.1145/3618315
[72] Zhijie Wang, Yuheng Huang, Da Song, Lei Ma, and Tianyi Zhang. 2024. PromptCharm: Text-to-Image Generation through Multi-modal Prompting and Refinement. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–21. https://doi.org/10.1145/3613904.3642803
[73] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems* 135 (2022), 364–381.
[74] Haijun Xia, Bruno Araujo, Tovi Grossman, and Daniel Wigdor. 2016. Object-oriented drawing. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 4610–4621.
[75] Haijun Xia, Ken Hinckley, Michel Pahud, Xiao Tu, and Bill Buxton. 2017. WritLarge: Ink Unleashed by Unified Scope, Action, & Zoom. 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, 3227–3240. https://doi.org/10.1145/3025453.3025664
[76] Zihan Yan, Chunxu Yang, Qihao Liang, and Xiang 'Anthony' Chen. 2023. XCreation: A Graph-based Crossmodal Generative Creativity Support Tool. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 48, 15 pages. https://doi.org/10.1145/3586183.3606826
[77] Ryan Yen and Jian Zhao. 2024. Memolet: Reifying the Reuse of User-AI Conversational Memories. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 1–22.
[78] 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 (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–21. https://doi.org/10.1145/3544548.3581388
[79] Chao Zhang, Cheng Yao, Jiayi Wu, Weijia Lin, Lijuan Liu, Ge Yan, and Fangtian Ying. 2022. StoryDrawer: A Child–AI Collaborative Drawing System to Support Children's Creative Visual Storytelling. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3491102.3501914
[80] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Paris, France, 3813–3824. https://doi.org/10.1109/ICCV51070.2023.00355