Brickify: Enabling Expressive Design Intent Specification through Direct Manipulation on Design Tokens

Xinyu Shi
School of Computer Science
University of Waterloo
Waterloo, ON, Canada
xinyu.shi@uwaterloo.ca

Yinghou Wang
Graduate School of Design
Harvard University
Cambridge, MA, United States
yinghouwang@gsd.harvard.edu

Ryan Rossi
Adobe Research
San Jose, CA, United States
ryrossi@adobe.com

Jian Zhao
School of Computer Science
University of Waterloo
Waterloo, ON, Canada
jianzhao@uwaterloo.ca

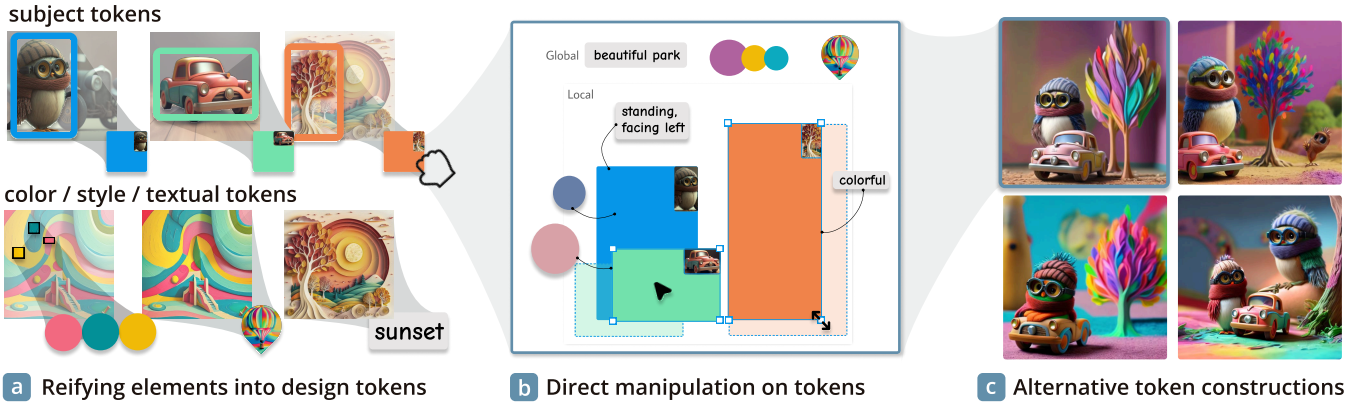

Figure 1: Brickify introduces a visual-centric interaction paradigm to specify design intent for controllable image generation: (a) users start with reference images and identify design elements (subjects, styles, colors, and concepts) which can be reified into interactive, reusable design tokens; (b) then directly manipulate on these tokens to build a visual lexicon to express how to construct these elements as a whole; (c) explore alternative compositions by reusing tokens and refining the visual lexicon.

ABSTRACT

Expressing design intent using natural language prompts requires designers to verbalize the ambiguous visual details concisely, which can be challenging or even impossible. To address this, we introduce Brickify, a *visual-centric* interaction paradigm — expressing design intent through *direct manipulation on design tokens*. Brickify extracts visual elements (e.g., subject, style, and color) from reference images and converts them into interactive and reusable design tokens that can be directly manipulated (e.g., resize, group, link, etc.) to form the *visual lexicon*. The lexicon reflects users' intent for both *what* visual elements are desired and *how* to construct them into a whole. We developed Brickify to demonstrate how AI models can interpret and execute the *visual lexicon* through an end-to-end pipeline. In a user study, experienced designers found Brickify more efficient and intuitive than text-based prompts, allowing them to describe visual details, explore alternatives, and refine complex designs with greater ease and control.

CCS CONCEPTS

• **Human-centered computing** → *Graphical user interfaces*; *Interactive systems and tools*; • **Applied computing** → *Arts and humanities*.

KEYWORDS

Design Intent Expression, Interaction Techniques, Direct Manipulation, Interactive Design Token

ACM Reference Format:
Xinyu Shi, Yinghou Wang, Ryan Rossi, and Jian Zhao. 2025. Brickify: Enabling Expressive Design Intent Specification through Direct Manipulation on Design Tokens. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3706598.3714087

CHI '25, April 26-May 1, 2025, Yokohama, Japan

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan*, https://doi.org/10.1145/3706598.3714087.

1 INTRODUCTION

Design is about the choices of visual elements (e.g., subject, style, color) and the construction of them towards an intended effect [25]; the decisions involved in this process are the art of design, embraced with a designer's creativity [36]. However, current text-to-image generation tools (e.g., DALL-E, Midjourney) shift design decision-making from designers to models, which are designed to *create for users* rather than *work with designers* [75, 83, 98]. Despite their ability to produce aesthetically appealing images for casual use, without designer's thoughtful planning, they lack the capability to create meaningful and professional design solutions that effectively convey intended visual messages [13, 17, 78]. One key barrier to keeping designers *in-the-loop* is communication [83, 98] — generative AI tools are designed to receive instructions *textually*, while designers often prefer to think and communicate *visually* [87, 96].

Consider the design case of crafting a Halloween poster: the designer collects some reference images (Figure 2a-2f), thinks about what visual elements to use, and after a while, forms a rough idea (Figure 2g) about how to compose them into a whole, and wants to work with generative AI tools for quick prototyping. However, the following three challenges arise.

First, clearly describing *what* visual elements to use in natural language is challenging due to inconsistent naming standards [39, 64]. For example, colors in Figure 2a might be described as *"light red, greenish teal, navy blue, with greyish black"* by some designers, and simply as *"orange, green, blue, and black"* by others. These subtle differences can lead to significant variations in hues and shades. Uploading reference images for model to "see" [17, 68] can help when the desired element dominates the image as in Figure 2d. However, in complex images as Figure 2c, specifying an exact element is harder. Designers might describe it as *"the abstract shape with curves in the middle right"*, but models often struggle with such ambiguous descriptions regarding the shape and position.

Further, precisely verbalizing *how* to construct those visual elements into a whole is hard. Deciding on relative scale and proximity is to relate the isolated elements with each other as interacting parts [25, 96]. However, both scale and proximity are continuous values, while language is often too discrete for fine-grained instructions. For example, describing Figure 2h as *"five pumpkins in front of a large building, with a moon above"* lacks precise size and position details, making it hard for the model to interpret their nuanced spatial relationship. The designer might finally obtain a reasonable version after multiple rounds of conversations with the tool, e.g., *"make the building a little bit larger"*, but this process is often time-consuming and tedious without guaranteed outcomes.

Lastly, the choices of elemental construction are infinite; however, prompting with texts limits the flexibility to *reuse* those visual elements to explore alternative constructions. Designers need to *copy-and-paste* an entire paragraph from previous prompts, then modify certain parts to change relationships or replace visual elements. Each iteration of exploration requires users to manually track where to change and where to keep among lengthy texts, which is tedious and error-prone. It is because existing generative AI tools treat each prompt as an independent request, without a mechanism to selectively separate persistent information (e.g., visual characteristics) from single-use prompts, making it inefficient to share information across multi-turn interactions.

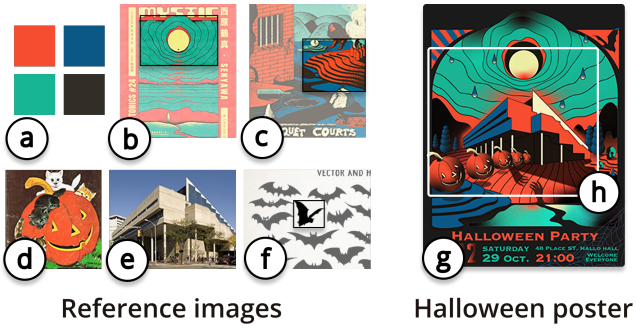

Figure 2: Design example of a Halloween party poster, showing (a) the color palette, (b-f) reference images with highlighted elements, and (g) the envisioned poster in designer's mind. (h) illustrates the spatial relationships between pumpkins, building, and moon. We have obtained the designer's consent to include this design in the paper.

To better align generative AI tools with designers' visual thinking process, we need a *visual-centric* interaction paradigm with versatile expressive power to facilitate graphic design that builds upon visual assets. We propose Brickify — specifying design intent via *interactive design tokens* that clearly carry the information of *what* primitive design elements to use (e.g., subject, color, style). Design tokens can be directly manipulated (e.g., drag-and-drop, resize, move, group, and link) to allow designers to precisely plan *how* the visual elements are constructed. The resultant construction — *visual lexicon* — can be translated into control signals for AI to faithfully generate the desired outcome, e.g., spatial layouts, relative scales, and the effect radius (e.g., applying a color to a specific subject or the entire image). Since design tokens can be persistent throughout the design process, users can efficiently *reuse* and recombine elements, avoiding the need to start from scratch for each iteration. We implemented and iteratively refined this interaction paradigm of Brickify in an interactive system named Brickify (Brickify refers to the paradigm while Brickify denotes the system).

We evaluated the interaction paradigm Brickify and its implementation in Brickify through an in-lab user study with 12 experienced designers. In a replication task simulating a scenario where designers have a clear intent, we compared Brickify with *textual-centric prompting*, finding that participants expressed design intent more precisely with less cognitive effort using Brickify and performed refinements faster, especially in complex designs. In an open-ended task with a simulated exploratory design scenario, designers found Brickify provided a controllable, expressive, and engaging design experience. Our findings offer insights for future research on designing interaction mediums for human-AI co-creation in broader visual design contexts.

2 RELATED WORK

AI-assisted graphic design involves *forming*, *specifying*, and *realizing* design intentions. We review how mood boards help form intentions, explore interactive techniques for specifying them, and summarize personalized image generation approaches to translate intentions into design outcomes.

2.1 Forming Intent from Reference Imagery

A common approach to develop design ideas is using references. Designers often start with exploring and organizing inspirations, then identify key design elements and strategically compose them [30].

Mood board usage. To support divergent thinking, designers often start with creating the mood board [26], a collection of images, shapes, colors, and other visual stimuli, as an aid to conduct visual research — framing, aligning, paradoxing, abstracting, and directing the design [58]. Mood boards are intentionally ambiguous, allowing different interpretations and serving as a tool for exploring creative possibilities [30]. How designers organize mood boards reflects their intended use, with studies exploring image arrangement in digital drawing and other fields [28, 43, 59]. Recent works [51–53] enhance mood boards with AI for semantic clustering, image recommendations, and material arrangement. Designers often decompose images into sub-elements [25] and then integrate them into cohesive designs [34]. Tools like MoodCubes [45] and MetaMap [47] help break down elements and enhance mood boards. Current generative AI tools allow users to upload reference images but lack flexibility in arranging or specifying sub-elements like on a mood board. Users cannot easily select which parts to use. Our approach integrates mood boards, enabling users to specify *what* sub-elements to use by converting them into design tokens.

Recombination of recognized elements. After identifying the elements of interest on the mood board, researchers have created a variety of tools to support the elements recombination. VisiBlends [16] and VisiFit [15] enable the blending of two semantic objects. Building on this, PopBlends [93] explores strategies for merging two conceptual keywords into pop culture images using large language models. Tools like 3DALL-E [57] and CreativeConnect [17] propose workflows for generating 3D designs and graphic sketches by suggesting keywords from design references and then refining them into detailed text prompts. Beyond keywords, Artinter [18] and GANCollage [92] enhance designer-client communication and style mixing by combining keywords and example imagery. Existing tools typically emphasize recombination outcomes under predefined rules for task-specific usage, such as blending by shape contours or mixing by styles. However, users have limited control over *how* to combine. This paper addresses this issue by enabling users to expressively construct the relationships among elements through direct manipulation of reified interactive tokens.

2.2 Specifying Intent by Interactive Strategies

The challenge of specifying design intent solely with natural language is well recognized [13, 80, 85, 98]. Subramonyam et al. [83] theorize how users translate their goals into clear intentions, highlighting the *instruction gap* where generative models are highly sensitive to language precision but human language tolerates variants in expression to communicate similar meaning. This challenge is evident across studies in various design domains, where casual users [60], graphic designers [55], manufacturing designers [35], and game professionals [90] struggle to articulate visual intent through the tedious process of prompt engineering. This friction of translating *visual* design into a *verbal* medium can be understood through the lens of design methodology literature [23, 86, 95]. Recent efforts to facilitate the intent expression fall into three categories: 1) decomposing the lengthy prompt into modular ones; 2) augmenting the text prompt with other modalities; and 3) resolving ambiguities in text prompts through direct manipulation [79].

Modular prompting. Managing lengthy prompts is challenging, and many research has explored to modularize them into manageable pieces. For instance, AI-chains [100] enables chaining individual prompts for end-to-end execution. In the similar spirit, ChainForge [4] supports comparing prompt variations between models. In visual design, Keyframer [88] uses "decomposed prompting" for step-by-step animation design. Similarly, Spellburst [3] and ComfyUI [22] employ node-based interfaces to modularize creative coding and image generation, integrating diverse control signals. Although these tools aid in flexibly writing and modifying prompts and allow viewing intermediate results, they still require users to specify their visual intentions in text, which does not fully address the precision of intent specification.

Multimodal prompting. Most recent commercialized tools (e.g., DALL·E3 [24], Adobe Firefly [31], MidJourney [62], Flux AI [2]) allow users upload images as global content and/or style references; however, users cannot add local annotations on the image to further illustrate their fine-grained intentions. Kaiber Superstudio [84] supports character-consistent generation by allowing users to specify a single local subject but does not support multiple subjects. Krea.ai [54] enables users to train on their own assets but requires multiple instances for each subject. Research efforts have also investigated different strategies for multimodal prompting. For example, DesignPrompt [68] allows users to input texts, images, and colors, then help translate them into a final textual prompt. However, translating visuals into text often leads to information loss, making it difficult to capture precise visual identities. PromptCharm [94] enables users to examine which part of the generated image corresponds to which part of the text prompt. Sarukkai et al. [74] introduce the coarse-to-fine sketch guided image generation. Although these tools are helpful in clarifying *what* visual elements to use to some degree, they fail to help users express *how* multiple visual elements relate with each other because they typically treat each input modality in isolation.

Prompting through direct manipulation. Some tools employ visual metaphors to help users specify their intentions through directly manipulating on visual objects rather than text prompts. For instance, TaleBrush [20] uses sketch lines to indicate narrative transitions, and PromptPaint [19] offers a paint-like interface for semantic prompt interpolations. Gestures are utilized to represent editing intentions, for example, drawing masks to inpaint [6], adding colored strokes to recolorize [103], and dragging points to edit pose or facial expressions [67]. While intuitive, these methods are task-specific, requiring users to re-learn interactions for each use case. Recent work extends this direction by aiming to integrate graphical user interfaces (GUI) with natural language interfaces

(NLI). DirectGPT [61] allows users to drag and drop graphical elements onto prompts, but these elements act as isolated symbols, losing their spatial relationships. DynaVis [89] combines NLIs with dynamically generated GUI widgets for visualization authoring, but requires users to specify the edit intention in text. Both DirectGPT and DynaVis address the ambiguity of continuous numerical expressions in textual prompts, yet they still force users to think textually and work primarily on texts.

2.3 Translating Intent into Design Outcomes through Computer Vision Techniques

Text-to-image generation is an active research topic since diffusion models emerging [42, 82], there are many amazing work has made the performance of text-to-image generation has reached a height that never been reached, such as DALL-E [11, 69, 70], GLIDE [65], Stable Diffusion [71], Imagen [73], etc. Research has also focused on improving controllability in image generation and editing.

Subject-driven and style-specific personalized generation. The task of *personalization*, introduced by Cohen et al. [21], aims to incorporate user-provided concepts absent from the training data into generated results. Early methods like Textual-Inversion [32] and DreamBooth [72] learn a *single subject* from *several user images*, while Custom Diffusion [56] and SVDiff [38] extend this to *multiple subjects* for recombination. Later, ELITE [97] and E4T-diffusion [33] make it possible to learn a *single subject* from *one image* but require the subject to be visually dominant. Break-A-Scene [5] allows learning *multiple subjects* from a *single image* with loose segmentation masks. Personalization also applies to styles: StyleDrop [81] fine-tunes models for style customization, while Style-Align [41] ensures style consistency without fine-tuning.

Spatial-aware controllable image generation and editing. To guide the large-scale pre-trained diffusion models to generate images following the spatially-localized conditions (e.g., depth map, segmentation, pose, etc.), Controlnet [102] embeds task-specific networks, while T2I-Adapter [63] offers a lighter adapter-based solution, and Composer [44] provides more flexible control with a larger model. For local image editing, initial methods [7, 8, 40] rely on precise text descriptions. Imagic [48] allows description-free editing but requires time-intensive fine-tuning, whereas Blended Latent Diffusion [6] enables faster editing without fine-tuning.

The advancements in the computer vision field are powerful and hold great potential. However, without intuitive interactions that allow users to precisely express their underlying intent, users cannot fully benefit from these capabilities. Our work focuses on innovating the user interaction, using these off-the-shelf methods as the technical foundation to realize our proposed *visual-centric* interaction paradigm.

3 ITERATIVE USER-CENTERED DESIGN

In collaboration with designers for about nine months, we employed an iterative design approach to define the *visual-centric* interaction paradigm, Brickify, and develop the Brickify system. The design process consists of four stages (noted as **S1-4**), with different participants involved in each stage. In this section, we introduce the participants and procedures for S1, S2, and S3, and discuss the findings and derived design goals from S1 (Section 3.1). To provide a holistic view, we briefly highlight the key design decisions made from S2 (Section 3.2) and S3 (Section 3.3), with further details provided in Section 4.

S1: Problem understanding (2 months) — interviews with six designers to identify challenges in using generative AI tools.
S2: Early Prototyping (4 months) — weekly co-design sessions with an expert designer to design the interaction paradigm and develop the early working prototype;
S3: Prototype Iteration (2 months) — informal testing involved six designers to collect feedback and iteratively refine the design;
S4: System Evaluation (1 month) — a user study with two tasks compares the visual-centric interaction paradigm of Brickify with the textual-centric one and examines how users interact with Brickify, which will be described in Section 7 and Section 8.

3.1 Problem Understanding: Interview Study

We conducted semi-structured interviews with six design experts (E1-6), with the aim to understand: (1) how designers approach prompting the generative AI models to craft graphic designs; and (2) what challenges they have encountered.

3.1.1 Participants and Procedure. Participants were recruited via email lists and social media, screened through a pre-test survey on design experience and familiarity with text-to-image generation tools. All had over two years of experience in graphic design, familiar with and regularly used the text-to-image generation tools in their work. Participants provided consent and were compensated with $20 for a 45-minute study session. We asked each participant to share at least one recent design project involving using text-to-image AI tools to reflect on how they use them.

3.1.2 Identified Challenges. We summarize the following challenges designers encountered when using generative AI tools.

C1: Failure to convey designers' attended elements to AI. Participants consistently began their design projects with visual research, using mood boards to collect inspirational images on a canvas in Figma (5/6) or Photoshop (1/6). They emphasized *"it's for understanding how pieces can fit together in my head."* -E3, aligning with prior studies [43, 52, 58]. Some participants (E1, E2, E4) grouped references spatially by element type, while others (E3, E5, E6) used annotations to mark elements. E5 explained, *"I'm not looking at the whole image, just the parts that matter to my design."* This selective focus, known as *active vision* [29, 37], is central to *visual thinking* [96]. While designers use *active vision* to pinpoint specific details such as a particular texture, a color theme, or the composition of shapes, AI models often lack the ability to recognize or prioritize these details in the same way. As E4 explained, *"The AI seems to understand the image globally, but I need it to work with specific parts."* Designers have to communicate the visual elements they are focusing on with AI through natural language, a medium that *"super hard for describing fine visual nuances."* -E1

C2: Difficulty in verbalizing element relationships. Participants emphasized the complexity of composing visual elements in a design, as E2 described, *"like constructing a house, you must place each brick properly."* E3 explained, *"It's not just about having the*

pieces; how to balance their weight is also important — some are more important, while others just for decorating." Designers must create focal points, balance hierarchy, and manage spatial placements, but articulating these relationships in words is difficult, especially when elements are intertwined. As E6 noted, *"planning that (how to compose them) in my head is hard enough, translating (the entire mental image) into a sentence feels much more difficult, I often sketch them down then describe."* This challenge often forces designers to simplify their ideas. As E2 mentioned, *"I tend to only ask for simple structures like centering a certain one (element)"*, but such compromise often results in outputs that *"lack the balance and spatial nuance we have been taught and always pursue in design."*

C3: Inefficiencies in iterative refinements. Designers face significant challenges in refining visuals through current text-to-image tools, as the conversations with AI is linear and lacks the mechanism to share key visual information in multi-turn dialogues. E4 pointed out, *"each change feels like starting over. I need a way to go back to tweak some [elements], but I don't want to touch certain ones I already feel good about."* As E2 explained, *"I have to copy and paste descriptions into every prompt, just to keep that part, but even so, it often get changed."* Without a way to selectively preserve certain visual information and design decisions, designers are forced to manage these iterations manually, making the design refinements inefficient and the creative flow disrupted.

3.1.3 Design Goals. To tackle the identified challenges, we articulate the following key design goals to drive the design of Brickify.

DG1: Support externalization of selective focus on primitive elements. Designers often focus on specific elements in reference images, but current systems require uploading entire images and verbally explaining their focus (C1). This creates a gap between what designers attend to and what the system processes. To bridge this gap, designers should be allowed to externalize their selective focus of elements into visual representations. This might involve allowing easy annotation, grouping, and flexible organization of these elements. By making these elements tangible, we aim to enable designers to interact with them directly, facilitating both their cognitive process and communication with AI at the element level.

DG2: Enable spatial management and visual communication of element relationships. Designers view elements as interdependent in a design, but articulating relationships such as scale, hierarchy, and spatial proximity is challenging (C2). As such, the user interaction should provide a flexible 2D workspace where designers can visually arrange and manipulate elements, defining relationships intuitively, and reducing reliance on verbal descriptions. The goal is to establish a shared visual structure that allows designers to clearly define the composition while enabling AI to accurately interpret and understand it, improving communication of complex elemental compositions.

DG3: Facilitate element reuse and iterative refinement. Designers struggle with the linear nature of current conversational text-to-image tools, which lack mechanisms for selectively preserving or refining elements across iterations (C3). To address this, designers should be able to easily reuse individual elements and partial configurations from previous versions to reduce repetitive manual work. With such a reusing mechanism, designers could explore different design variations more efficiently.

3.2 Early Prototyping: Co-designing with a Designer

3.2.1 Procedure. In the early stages of our project, we engaged in a four-month co-design process with an expert designer, who has over eight years of graphic design experience. We held weekly 30-minute design meetings. During this phase, we collaboratively created early low-fidelity mock-ups using sketches and iteratively built non-functional prototypes in Figma. This collaboration focused on defining core components of Brickify and basic features in Brickify. Given her commitment, we include her as a coauthor.

3.2.2 Design Outcomes. We defined two key aspects of the Brickify interaction paradigm: 1) reifying [10] design elements into *design tokens*; and 2) enabling direct manipulation [79] on tokens, constructing the *visual lexicon*, to specify relationships. We identified two types of design tokens: *visual* and *textual*. Among the visual tokens, we included three core elements: subject, style, and color. We also defined five essential manipulation capabilities: *drag-and-drop*, *move*, *resize*, *group*, and *link* tokens. The interface operationalizing this paradigm was structured into three panels: 1) a mood board panel for organizing reference images and creating tokens, 2) a token manipulation panel for building relationships, and 3) a history panel to track versions. These design decisions led to the development of an initial working prototype. Details will be described in Section 4.

3.3 Prototype Iteration: Involving Another Six Designers

3.3.1 Participants and Procedure. To refine the initial design of Brickify and the early prototype, we recruited six additional designers with over one year of graphic design experience, each having used at least one text-to-image generation tool more than five times in the past three months. We began by walking them through the initial prototype and explaining how to interact with the system. Using our prepared reference images, we asked them to explore the system and generate multiple designs. Designers used a think-aloud method to inform us when they encountered difficulties or desired alternative functionality. At the end of the session, participants provided feedback and discussed their overall experience and suggestions for improvement. The entire study lasted about 45 minutes, and participants were compensated with the equivalent of $20 CAD.

3.3.2 Feedback and Design Refinements. Designers identified several inadequacies in the current Brickify design and suggested immediate feedback. Based on their suggestions, we strengthened the visual association between design tokens and their original imagery to improve clarity. Additionally, we introduced a *cross-referencing* feature to allow for more effective descriptions of relationships between subject tokens. Designers also expressed the need to accommodate both concrete instructions and abstract imagination, so we added an *imaginative token* to the interaction vocabulary. These refinements helped make the Brickify more expressive. Details will be described in Section 4.

4 BRICKIFY: A VISUAL-CENTRIC INTERACTION PARADIGM

In this section, we explain the design decisions and rationale behind Brickify, a *visual-centric* interaction paradigm that enables users to express design intent through *direct manipulation on design tokens*.

4.1 Design Tokens: Specifying What Elements to Use

We introduce *design tokens* as the externalizations of designers' attended design elements (DG1), reifying [10] the abstract visual information into concrete first-class graphical objects that can be directly manipulated and reused.

4.1.1 Token types: Being polymorphic to ensure expressiveness and extensibility. A key design insight in Brickify is that *all types of design elements and intentions should be regarded as tokens*. The goal is not to support a complete set of all possible design elements but to build an *extensible* paradigm with the affordance to accommodate different types. Such *polymorphism* [10] is essential for maintaining a simple interface with consistent interaction logic. During the early prototyping (S2), the designer expressed the desire for precise control over style, colors, and subject identity. Later in the prototype iteration stage (S3), participants added that they also appreciated the model's hallucinations for certain details. For instance, one participant noted, *"I will leave it to the model to decide how an exact 'joyful' facial expression looks like."* Thus, we categorize design tokens into three types: *visual, textual,* and *imaginative* (Fig. 3).

> **Visual token** carries the visual information such as the subject, style, and color, reified from reference images.
> **Textual token** complements visual tokens by conveying information that is easier to express through language, such as adjectives for emotions or verbs for gestures.
> **Imaginative token** mediates the initiative between designers and models, indicating *where* the model should intervene and *how much* imagination is needed.

4.1.2 Token appearances: Balancing fidelity with re-envisioning potential. The design tokens can be regarded as a visual abstraction depicting the elements graphically. When designing their appearance, we balanced between *fidelity* and *re-envisioning potential*. Tokens need to be visually distinct, allowing users to easily identify what element they represent while retaining enough abstraction for designers to re-imagine them in new contexts.

In early co-design sessions (S2), we used geometric shapes to represent different elements, e.g., rectangles for subjects, circles for colors, and filled rectangles with different colors to distinguish between subjects. Hovering over a token would highlight the original source in the reference image. However, in the later prototype iteration stage (S3), we found that as the number of subjects grew and design complexity increased, participants struggled to track which token represented which subject, frequently switching between the mood board and token manipulation panels to confirm identities. To address this, we refined the subject tokens by attaching a small cropped image of the subject to the token's corner. For style tokens, we represented them by transferring the style to a standard image.

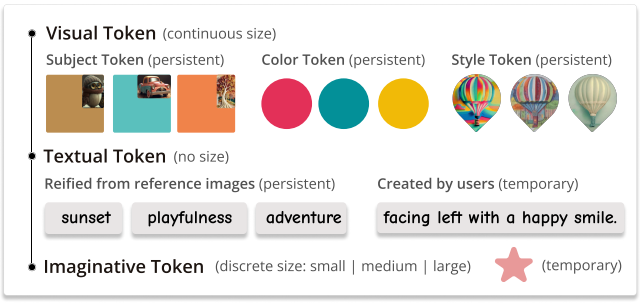

Figure 3: The definition of design tokens in Brickify: visual, textual, and imaginative tokens. Each type of tokens has their own appearances and life-cycles.

4.1.3 Token life-cycles: Offering both persistent ones for reuse and temporary ones to avoid overwhelm. Towards facilitating element reuse and iterative refinement (DG3), we make a distinction between *persistent* and *temporary* tokens. Persistent tokens are used for core content elements that are repeatedly referenced throughout the design process, while temporary tokens represent contextual details or single-use modifications.

Designers typically plan a design by considering content, construction, and context. For instance, *"two girls* (content) *next to each other in the center of the image* (construction) *are dancing surrounded by flowers* (context)*"*. Designers often explore alternative design possibilities by altering the construction or context while keeping the same visual elements. To support this process, content-related tokens are persistent for *reuse* across different design variations, while context-related tokens remain temporary to avoid clutter. We implemented this distinction by separating token creation into two panels: tokens created in the mood board panel are persistent, with each use being a copy of the original, ensuring the flexibility of reuse. In contrast, tokens in the manipulation panel are temporary and deleted when the panel is cleared to prevent interface clutter.

4.2 Direct Manipulation on Tokens: Expressing How to Construct Elements

Direct manipulation [79] has long been integral to designers' workflows, especially for rapid prototyping and visual planning. In Brickify, users express *how* they want to construct elements through direct manipulations on design tokens (DG2).

4.2.1 Intuitive actions to reflect intentions. The mapping between intention and action should be intuitive, allowing users to engage through *technical reasoning* [66] rather than *procedural learning*. In collaboration with designers, we defined the following actions to reflect design intent, constructing the *visual lexicon* by manipulating on tokens, as shown in Fig. 4.

> **Drag-and-drop.** Users create persistent tokens (subject, style, color, concept) from reference images in the mood board panel. To use these tokens, they can drag their copies and drop them onto the token manipulation panel, where each token can be reused multiple times without limits.
> **Move.** Users can freely move subject tokens to define the spatial relationships between subjects. While other tokens

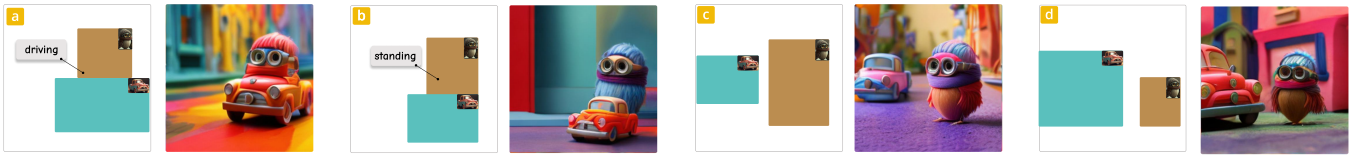

Figure 4: Demonstration of exploring different compositions through direct manipulation on design tokens. (a)–(d) show how adjusting sizes and positions of the owl and car tokens changes their relationships in the outcomes.

(e.g., color, style, textual) can also be moved, the movement of these tokens is purely for organizational purposes without encoding spatial relationships, as their function is to modify or describe attributes of the subject.

Resize. Resizing tokens adjust their scale. Color tokens indicate proportional weights (e.g., primary vs. secondary colors), while style tokens work similarly. Resizing subject tokens specifies their sizes in the output, and resizing imaginative tokens controls the extent of AI-imagined details. Textual tokens cannot be resized.

Group. Grouping tokens helps manage multiple elements easily. For example, users can group 3-5 colors into a color theme or apply several colors to a single subject.

Link. Design elements are often interconnected. For example, a color token can be linked to a specific subject, applying only to that subject, or left unlinked to apply globally. Links specify the relationships between tokens, such as binding colors or textual descriptions to subjects.

Cross-reference. Subject tokens often reference one another to specify certain relationships. For example, in the phrase *"an owl is driving the car"*, describing the owl's behavior using a textual token requires referencing the car's token. To cross-reference, users can assign a name to a token to refer and tag the name in a textual token.

4.2.2 Flexible action reuse. In addition to reusing tokens, designers should be able to reuse their previous actions to explore alternative design paths (DG3). Since actions are reflected in the construction of design tokens, the visual lexicon, we support action reuse by recording each lexicon created. Designers can refine their work based on this visual lexicon rather than redoing previous actions, enabling more efficient iteration and exploration.

4.2.3 Intermediate action outcomes. During prototype refinement (S3), designers expressed the need for immediate feedback on their actions to assess how design tokens respond and evaluate the results. However, the current generative models have noticeable inference times and high computational costs, making instant feedback for every action impractical. To address this, we introduced feedback at a higher granularity. Since the execution process follows a sequence — first composing the layout, then aligning the style, and finally applying colors — we provide intermediate results at each step. As model inference times improve, we envision the possibility of real-time feedback for more responsive interaction.

5 USAGE SCENARIO

Before diving into the design process and the implementation in detail, we walk readers through an example usage scenario to express design intent through Brickify. Imagine Stella, a designer, working on a children's storybook about the adventures of an owl. The story follows the owl as he travels in different landscapes with his trusty car and friend.

Creating design tokens. Stella begins by gathering inspirational images to define the look of the characters and the feel of the scenes and importing them to the Mood Board Panel (Fig. 5A). She draws bounding boxes around the owl, car, and tree to create their subject tokens. She then creates color and style tokens with the corresponding tools shown in (Fig. 5a). To set the thematic tone, she adds a textual token for "playfulness".

Manipulating design tokens. She drags and drops (Fig. 5(b1)) the created tokens to the Token Manipulation Panel (Fig. 5B) to build her story (Fig. 5(b2–b6)). She imagines the owl parking under a tree, waiting for his friend. She resizes and positions the owl, car, and tree tokens to define their spatial relationships, then links textual tokens to specify the owl standing behind the car and facing left. For the tree, which she imagines as "colorful" but undefined, she links an imaginative token to let the AI decide it. Stella then groups the color tokens for a cohesive theme and adds a style token. For the background, she creates a textual token of "beautiful park".

Reusing design tokens. In the next scene, where the owl's friend joins, Stella reuses parts of the previous visual lexicon, making slight adjustments and dragging and dropping another subject token of owl from the Mood Board Panel as his friend. By reusing design tokens, she streamlines her workflow and avoids redundant work. All generated results and their visual lexicons are organized in the History Panel (Fig. 5C).

6 BRICKIFY SYSTEM IMPLEMENTATION

In this section, we explain how Brickify extracts primitive design elements from reference images to create design tokens (Section 6.1) and transforms the tokens together with users' actions on tokens into control signals for models to process (Section 6.2).

6.1 Design Token Creation

6.1.1 Subject token. Users create a subject token by drawing a bounding box around the desired subject using the subject tool. To ensure that generative models accurately capture the visual details specified by users, we employ the `SAM` [49] model to extract segmentation maps and then use the `Break-A-Scene` [5] approach to fine-tune the `Stable Diffusion (v2.1)` model. This process learns the subject's visual identity and binds each subject to a specific token within the model for later use. To accommodate multiple reference images, we concatenate them into one image before fine-tuning because `Break-A-Scene` can only learn subjects within one single image.

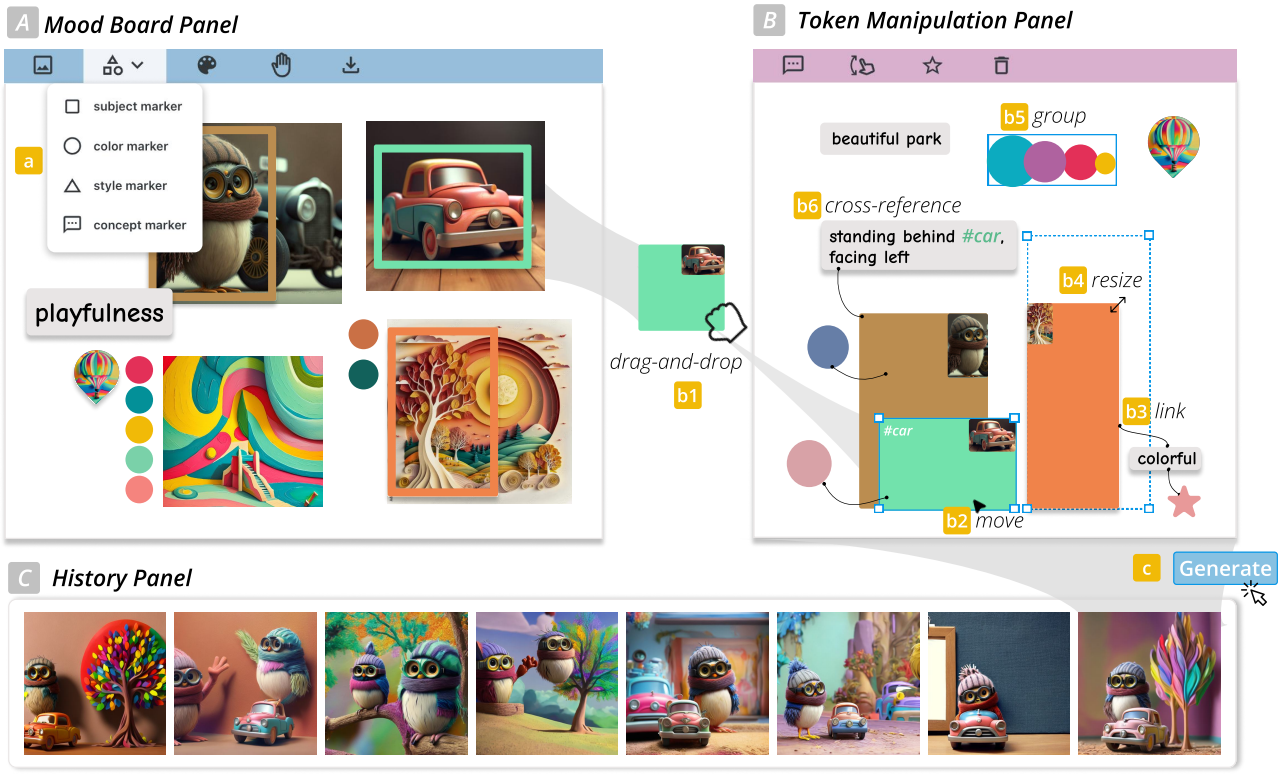

Figure 5: User interface of Brickify, consisting of three panels: (A) *Mood Board Panel* for arranging reference images and creating persistent design tokens (subject, color, style, concept), which can be drag-and-dropped (b1) into (B) *Token Manipulation Panel* for direct manipulation (b2 – b6). Clicking the *Generate* button (c), generated results are organized in (C) *History Panel*.

6.1.2 Color token. Users can extract color tokens both automatically and manually. Using the color tool and clicking on an image automatically extracts five dominant colors as color tokens based on `K-Means` clustering. If the extracted colors do not meet the user's needs, they can click on them to manually change their colors with the color picker. If they want to create more color tokens based on one image, they can click again on the image to create a circle with a random color, allowing users to manually select a color with an eyedropper tool to create a color token.

6.1.3 Style token. Users indicate their desire to use an image's style by clicking on it with the style tool. We leverage `Style-Align` [41] to transfer the image's style to a standard balloon image, which is then cropped to the marker shape, creating a style token.

6.1.4 Concept token. Concept tokens capture the high-level spirit or emotional feeling of an image in textual format. When users click on an image with the concept tool, `GPT-4o` describes the feeling and atmosphere of an image, summarizing it into five keywords.

6.2 Visual Lexicon Execution

To ensure a smooth iterative design experience, we selected the approaches for the visual lexicon execution that do not require training or fine-tuning on diffusion models. It should be noted that the field of computer vision evolves rapidly and methods could be replaced as better solutions emerge, our goal is to provide a feasible technical pipeline for executing the visual lexicon users create. The execution consists of four primary steps: handling layout, style, global colors, and local colors (Fig.6). This order is deliberately designed to prevent visual effects from being overridden. For example, handling the style inevitably changes the colors to some degree, so color adjustments must come afterward. Similarly, global colors are handled before local colors.

6.2.1 Extend keywords description. Users often use only keywords in textual tokens, but diffusion models perform better with complete sentences as prompts. We thus extend textual tokens into sentences using `GPT-4o`. The size of the imagination token determines the level of detail added, with three levels: small, medium, and large. If no imagination tokens are used, only factual information is stated without any added imagination.

6.2.2 Compose the layout. To compose subjects into the desired layout, we use `BoxDiff` [101], which constrains image generation with spatial control guidance. For the foreground, it takes the subject token placements as bounding boxes and related textual keywords describing each bounding box as input, triggering the special tokens in the pre-trained diffusion model fine-tuned by `Break-A-Scene` to generate subjects according to the specified layout. It also handles the background generation with the given text description.

Figure 6: The technical pipeline of Brickify interprets and executes the visual lexicon step-by-step, using off-the-shelf methods.

6.2.3 Align the global style. By default, the image composed in the layout at the previous step is in a realistic, natural style. To align it with the user-specified style token, we use the `Style-Align` [41] method. `Style-Align` performs shared self-attention with the reference style image at each diffusion denoising step to achieve style alignment during the image reconstruction process.

6.2.4 Apply global colors. To apply the global color palette to the image, we employ the `HistoGan` [1] to recolorize images based on the given palette. Specifically, we use the pre-trained checkpoint of `Universal model-0`. This method is based on generative adversarial networks (GANs) and optimizes the recolored image to match the color proportion distribution by projecting color histogram features into the model's latent space.

6.2.5 Calibrate local colors. Lastly, we handle the color tokens attached to local subjects. `Blended Latent Diffusion` [6] is leveraged to perform local color modifications on the image. This method targets the local editing of generic images, where the desired edits are confined to a user-provided mask without touching the rest.

7 USER STUDY DESIGN

We conducted an in-lab user study to evaluate the effectiveness of Brickify in intent expression and users' experience when interacting with the system of Brickify. The study involved two tasks with the same set of participants.

7.1 Participants

We recruited 12 experienced designers via social media and mailing lists. All participants hold a formal design degree and have more than 3 years of experience in graphic design. Participants rated their frequency of using generative AI tools for text-to-image generation on a 5-point Likert scale (1 = "never" to 5 = "very often"), with an average rating of 3.33. Their detailed demographic information is listed in Appendix 1. Participants took part in the study remotely. All study sessions were audio and video recorded. The entire study lasted about 75 - 100 minutes, and participants were compensated with the equivalent of $30 CAD. The study was approved by the university's ethics review board.

7.2 Study 1: Interaction Paradigm Comparison

Study 1 uses a replication task to simulate a scenario where designers have a well-developed idea in mind. The goal is to answer the research question (**RQ1**): How does the *visual-centric* interaction paradigm of Brickify compare to the *textual-centric* paradigm in terms of clarity, mental effort, and time investment for expressing design intent?

Figure 7: Reference and target images for Study 1. For each condition (Easy and Hard), users work with two versions: one created from scratch and a second tweaked version where users adjust their original description to match the modified target image.

7.2.1 Experimental Design. We use a 2 × 2 within-subject design with two primary factors: Technique (Brickify or Baseline) × Difficulty (Easy or Hard). The Brickify condition, as described in Section 4, is our proposed interaction paradigm, accessed through the Brickify interface shown in Figure 5, with the *Generate* button disabled. In the Baseline condition, participants describe their design intentions by typing textual prompts in a Google Doc, with the ability to refer to reference images by their provided names. We avoided using existing commercial text-to-image interfaces like Dall-E or MidJourney as baselines due to their differing interaction designs — Dall-E allows uploading an unlimited number of images but lacks a clear mechanism to reference them while MidJourney uses command-style prompts and requires images to be in URL format. To ensure fairness for participants familiar with different tools, we thus provided this general textual-centric prompting method reflecting common generative model interactions as a baseline.

The Difficulty levels (Easy and Hard) are determined based on the number of visual elements and the complexity of their compositions. To establish these conditions, an expert designer selected elements from a set of reference images and created initial versions for both difficulty levels. For the Easy condition, the design involved fewer elements and simpler compositions, while the Hard condition included more elements with more intricate arrangements. The designer then further refined the compositions to create a second version for each condition, as shown in Fig. 7.

Participants were asked to express their design intent for the target images across the two Difficulty levels (Easy and Hard) using two interaction paradigms: Brickify (visual-centric) and Baseline (textual-centric). For each Difficulty level, participants first described the design of a target image (Easy-v1 or Hard-v1),

then refined their expression to produce a modified version (Easy-v2 or Hard-v2), simulating a real-world design refinement process.

7.2.2 Measurements. To assess whether user expressions sufficiently described the target image, we adopted a human-evaluation approach by recruiting three external raters to assess all participants' expressions under both Techniques. Raters self-reported frequent use of text-to-image generation tools ($M = 4.33, SD = 0.58$; scale: 1 = "never" to 5 = "very often") and demonstrated being knowledgeable in prompt engineering ($M = 3.67, SD = 0.58$; scale: 1 = "no experience" to 5 = "expert"). The three raters performed the evaluation independently on five 7-point Likert Items (i.e., element coverage, size clarity, position clarity, style clarity, and color clarity) followed the predefined rubric (see Appendix A.3). The expressions were order-randomized for each rater. The rating process took around 2 hours and the raters received $50 CAD for their time.

We chose not to use AI models to directly execute expressions and generate final outcomes to compare due to the following reasons. First, there is no off-the-shelf techniques built on top of Stable Diffusion 2.1 (the same base model in our technique) that resemble Baseline — taking multiple references as input and leveraging their subjects and styles — to make it comparable. Second, as in this study we focus on the expressivity aspect, uncertainty and complexity in the process of execution for AI models may introduce compounding factors unrelated to the quality of user expressions. Therefore, we instead rely on human raters, whose evaluations could more effectively reflect the quality of the expressions from the message receiver perspective.

To this end, the measurements for Study 1 included 1) participants' responses to five questions evaluating intent expression, 2) external ratings for participants' expression in different conditions, 3) task completion times for both the initial and refined design versions, 4) participants' preferences between the two Techniques, and 5) self-reported cognitive load during the tasks.

7.3 Study 2: Brickify Exploration

Study 2 is an open-ended task, without comparison with other systems, designed to explore the research question (**RQ2**): How does Brickify influence users' creative exploration when they start without a clear intent?

Figure 8: Reference images used in Study 2.

7.3.1 Task design. In this task, participants assumed the role of junior graphic designers tasked with creating a graphic series for a children's storybook about the adventures of an owl. The senior designer provided four reference images (Figure 8) to define the visual characteristics. Participants were asked to create three images depicting scenes where the owl, with or without his friend and car, embarks on an adventure. The task required maintaining visual consistency across all images. There was no time limit, and participants worked until they felt their designs were complete.

7.4 Procedure

After signing the consent form, participants were given an overview of the study procedure, duration, and data collection details. The studies were conducted remotely via Zoom, with participants accessing Brickify through the web browser. A brief training session, including a toy example, was provided to demonstrate the use of Brickify and explain the text-based prompting in the Baseline condition. Participants could familiarize themselves with the tools before starting the tasks.

Participants began with Study 1, completing both Easy and Hard tasks using Brickify and the Baseline, with no time limits. The sequence of the four trials (Baseline-Easy, Baseline-Hard, Brickify-Easy, Brickify-Hard) was randomly assigned across participants, as a full counter-balancing was not feasible. After each trial, participants filled out questionnaires to rate their intent expression experience. Upon completing Study 1, they filled out the questionnaire on their preference and cognitive load on a 7-point Likert Scale. Next, participants proceeded to Study 2. After completing this task, they rated their experience with Brickify using a post-study questionnaire for Creativity Support Index (CSI) [14]. A semi-structured interview was conducted to gather feedback on participants' experiences with Brickify. Participants were encouraged to share general comments on any aspect of the study and then prompted on specific aspects, including interface usability frustrations, challenging intention expression cases, prior difficulties with text-centric GenAI tools, and whether similar issues arose with Brickify. They were also asked about the system's impact on their approach to solve design problems, exploration of design options, and suggestions for improvement. Observations noted by the experimenter during the session were also discussed.

7.5 Data Analysis

To analyze the qualitative feedback, we analyzed interviews using thematic analysis, employing both inductive and deductive approaches. Two researchers collaboratively analyzed and open-coded the transcribed interviewees' responses, employing affinity diagramming to sort the initial codes onto cards. Then, they discussed and reconciled any discrepancies in the coding process to ensure a consistent and accurate representation of participants' perspectives. Through iterative discussions and the organization of these codes, we identified a number of recurring patterns and themes within the interview data.

8 USER STUDY RESULTS

8.1 Self-Rated Design Intent Expression

We conducted the non-parametric *Aligned Rank Transform (ART) ANOVA* [99] statistical analysis on the ordinal Likert-Scale subjective ratings for Study 1 to understand the influence of Technique and Difficulty on users' self-reported design intent expression experience (Q1-Q5). Results show that there is a significant main effect of Technique on users' self-reported design intent expression experience across all five questions, while Difficulty had significant effects for Q2, Q3, Q4, and Q5 but not for Q1. The interaction effect between Technique × Difficulty was significant for Q3 only. The post-hoc multi-factor contrast tests following the

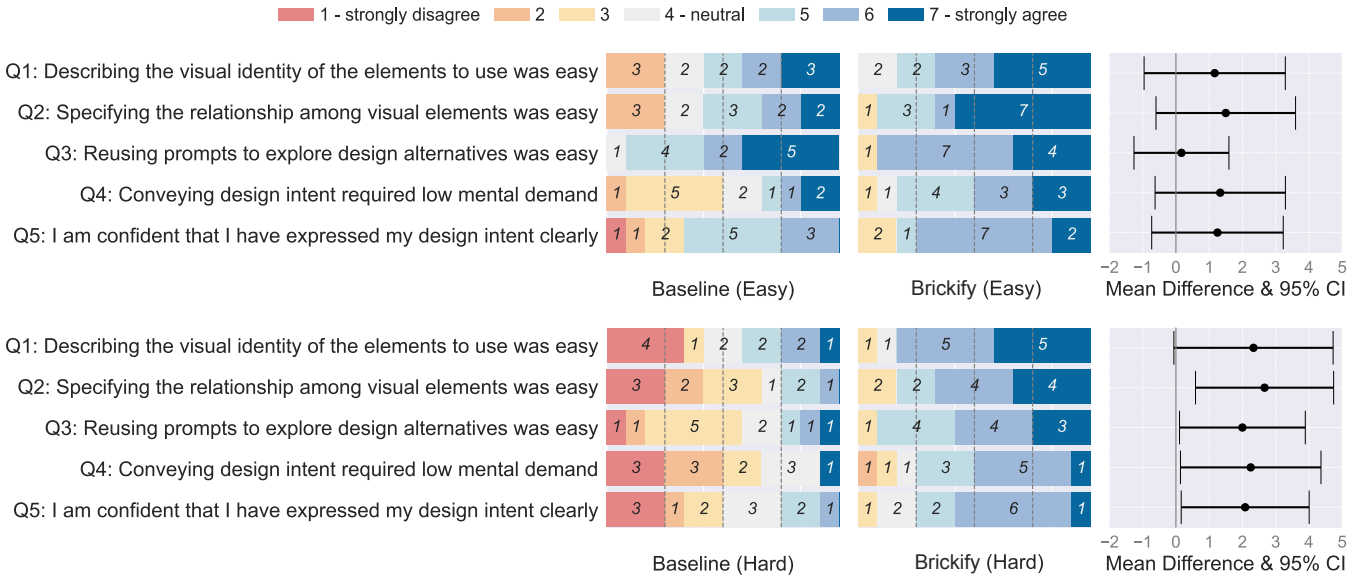

Figure 9: Participants' response for Study 1 when rating the 7-point statements for Baseline and Brickify interaction paradigm under Easy and Hard conditions. Dots are the mean differences of Brickify compared to Baseline.

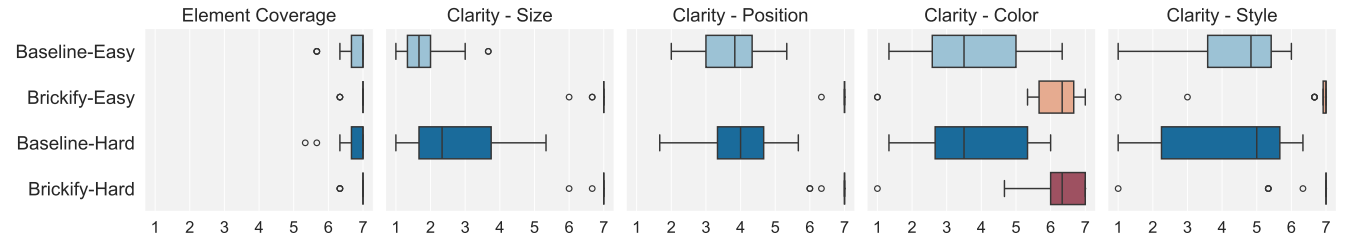

Figure 10: External rating for the quality of expressions that participants produced in Baseline and Brickify regarding the element coverage, the clarity of size, position, style, and color.

ART-C [27] procedure were conducted to identify the exact differences. The results show that participants' ratings on Q2, Q4, Q5 were significantly higher for Brickify than the Baseline in the Easy task. For Hard task, participants' ratings for Brickify on all five questions (Q1-Q5) were significantly higher than the Baseline.

The analysis, along with the detailed scores in Fig. 9, demonstrates that Brickify is more effective than the text-centric interaction paradigm in supporting users' design intent expression, particularly in *higher difficulty* tasks. Brickify simplifies describing visual identity, specifying relationships, reusing prompts for alternatives, reduces cognitive load, increases expression confidence, and enhances the overall design expression experience.

8.2 External-Rated Expression Quality

Figure 10 shows the external ratings for participants' expressions across five items — element coverage, clarity of size, position, color, and style (the rubric is shown in Appendix A.3). The reliability of the ratings was measured using a two-way random Intraclass Correlation Coefficient (ICC). The ICC values for each item ranged from 0.688 to 0.930, with an average of 0.817, indicating acceptable reliability. We use the average score from the three raters for each expression for further statistical analysis. Across different Difficulty levels, participants in both the Baseline and Brickify conditions successfully covered most elements in the target images. While Brickify received slightly higher scores for element coverage, the difference was not statistically significant. For the perceived clarity, the two-way ANOVAs indicated that the choice of Technique (i.e., Baseline vs. Brickify) is the primary factor significantly influencing the clarity of size ($F_{1,92} = 892.15$, $p < .001$), position ($F_{1,92} = 501.02$, $p < .001$), color ($F_{1,92} = 58.14$, $p < .001$), and style ($F_{1,92} = 51.80$, $p < .001$) regardless of task difficulty. These results suggest Brickify provides a more effective approach for reducing the ambiguity in intent expression.

8.3 Task Completion Time

In Study 1, on average, participants took longer to complete the initial version in the Brickify condition compared to the Baseline condition ($\bar{t}_{\text{Baseline\_Easy}} = 324s$ vs. $\bar{t}_{\text{Brickify\_Easy}} = 474s$, $\bar{t}_{\text{Baseline\_Hard}} = 444s$ vs. $\bar{t}_{\text{Brickify\_Hard}} = 465s$), detailed data is shown in Figure 11. However, the differences were not statistically significant. An ANOVA analysis of the model Initial Completion Time ~ Technique × Difficulty revealed no significant

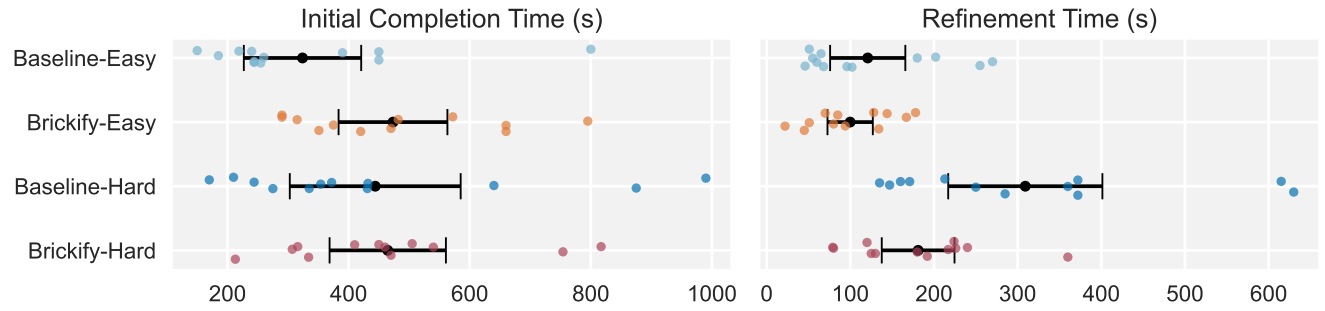

Figure 11: Users' Initial Completion Time (left) and the Refinement Time (right) with Baseline and Brickify interaction paradigm under Easy and Hard conditions in Study 1. Black dots are means, bars are 95%CIs.

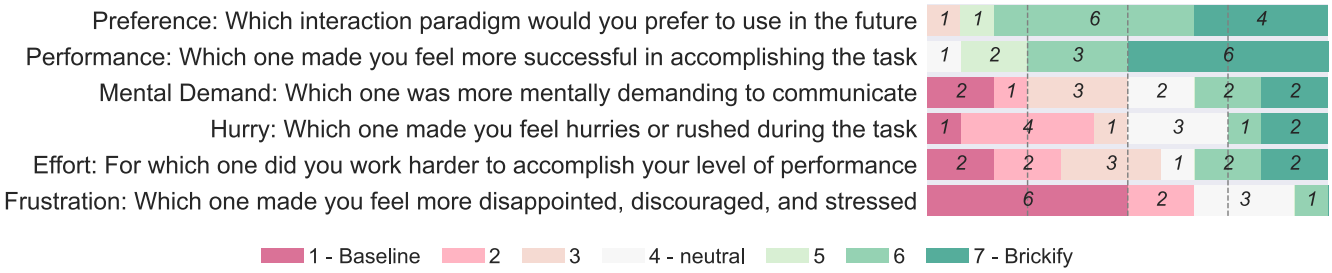

Figure 12: Participants self-reported preference and cognitive load that directly compare Baseline and Brickify.

main effects of Technique ($F_{1,44} = 2.93$, $p > .05$) or Difficulty ($F_{1,44} = 1.44$, $p > .05$), and no significant interaction between Technique and Difficulty ($F_{1,44} = 1.32$, $p > .05$). This result aligns with our expectations, as Brickify inherently involves many operations (e.g., creating, dragging, and constructing tokens) that participants have to perform to express their intent from scratch. The breakdown of Initial Completion Time shows that participants spent approximately one-third ($\bar{t}_{\text{Easy}} = 139s$, $\bar{t}_{\text{Hard}} = 165s$) of the total time in creating tokens and dragging them to the manipulation panel. The remaining time was dedicated to manipulating these tokens to construct the lexicon.

When entering the refinement stage, participants, on average, took less time in Brickify condition to refine their initial expression to achieve the modified version than Baseline condition ($\bar{t}_{\text{Baseline\_Easy}} = 121s$ vs. $\bar{t}_{\text{Brickify\_Easy}} = 99s$, $\bar{t}_{\text{Baseline\_Hard}} = 309s$ vs. $\bar{t}_{\text{Brickify\_Hard}} = 181s$). The ANOVA in the model Refinement Time ~ Technique × Difficulty shows significant main effects of Technique ($F_{1,44} = 5.81$, $p < .05$) and Difficulty ($F_{1,44} = 18.03$, $p < .001$), but no significant Technique × Difficulty interaction ($F_{1,44} = 2.97$, $p > .05$) on Refinement Time. Post-hoc Tukey HSD tests further show that users spent significantly less Refinement Time in Brickify on average by 128 seconds ($p = .027$) in Hard condition, while no significant difference found for Easy condition. This result can be attributed to the fact that participants did not create new tokens in Brickify during the refinement stage; instead, they focused solely on manipulating existing tokens. The faster Refinement Time highlights Brickify's strength in enabling quicker and more efficient modifications once the initial design intent is established.

8.4 User Preference and Cognitive Load

After completing Study 1, participants were asked to rate their preference between Brickify and the Baseline condition, as well as their cognitive load for each condition (Fig. 12). Participants (11/12) showed a clear preference for Brickify over the Baseline. On a 7-point Likert scale (1 = strongly prefer Baseline, 7 = strongly prefer Brickify), the mean preference rating of 6.0 was significantly above the neutral midpoint of 4. This indicates a strong preference for Brickify when participants had a clear design intent to express.

Regarding cognitive load, most participants (11/12) felt they were more successful using Brickify. Additionally, 7 out of 12 participants reported making less effort, and 8 out of 12 felt less frustrated, suggesting that Brickify simplifies the design intention expression and reduces users' frustration. However, only half of the participants felt reduced mental demand and hurry compared to the Baseline. As users engage more deeply in articulating their design intent, they may invest more time and mental effort in decisions such as token placements. Overall, Brickify enhances the design intent expression clarity and reduces frustration but still requires a certain level of mental effort and time commitment to fully engage with design token manipulations.

8.5 User Behavior

8.5.1 Token Usage. Figure 13 illustrates the number of participants using each token type across studies. Subject tokens were consistently adopted by all participants across all conditions, while color, style, and textual tokens were used by most participants (more than 8). There was a slightly reduced usage (but still more than half of the participants) in Study 2 for these tokens, because they focused

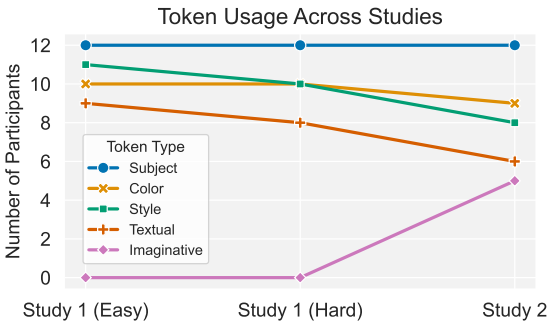

Figure 13: Token usage in Brickify across studies, indicating the number of participants have used each token type.

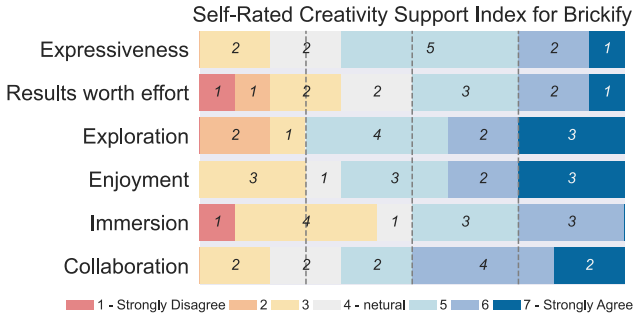

Figure 14: Self-reported Creativity Support Index for Brickify after Study 2.

more on shaping the narrative by manipulating subject tokens. Notably, no participants used imaginative tokens in Study 1, which may be because of its well-defined scenario with clear targets relied entirely on reference images. During Study 2, where the design scenario was more exploratory, five participants employed imaginative tokens for (1) enriching background descriptions (P2, P4, P10), (2) modifying subject tokens to adjust visual identity and/or actions, e.g., differentiating the owl's friend from the owl (P11), and (3) expanding text descriptions for global styles (P4, P9).

8.5.2 User Strategy. Across both studies, we observed several strategies shared across participants when interacting with Brickify to construct the visual lexicon. All participants adopted an on-demand approach for token creation, using tokens as needed rather than creating all tokens upfront. This led to frequent navigation between the mood board panel and the token manipulation panel. When constructing the visual lexicon, 9 out of 12 participants prioritized creating and positioning subjects first, followed by adding local colors (if any), and then applying global styles and/or colors. The remaining participants began by considering global styles and colors before adding subjects and their local colors. We also observed that most participants (N = 10) started with the background or underlying layers and worked progressively toward the foreground. As P2 explained in the interview, this approach likely stems from their unconscious habit of working with layer-based logic in tools like Adobe Photoshop and Illustrator: *"I always build layer by layer, back to front in Illustrator and it feels natural to follow that order."*

8.6 Self-reported Creativity Support Index

We utilize the Creativity Support Index (CSI) to measure the degree of creativity support for Brickify in the Study 2. Since Study 2 does not include a baseline for comparison, we present this self-reported rating as a reference point to better understand users' experiences, rather than drawing definitive conclusions. Participants rated six creativity support factors on a scale from 1 (strongly disagree) to 7 (strongly agree) shown in Fig. 14: expressiveness, results-worth-effort, exploration, enjoyment, immersion, and collaboration. Overall, Brickify shows strong support for creativity, effectively supports idea exploration, and is generally enjoyable to use. Although most participants rated positively, one participant strongly disagreed with the immersion factor, which may reflect interface limitations, as noted in an interview where the user suggested frequently used tools to create visual tokens should be directly available in the toolbar rather than in a list.

The factor of *results worth effort* reflects how effectively the model executes users' intent. Half of the participants rated it positively, while the rest were less satisfied. This disparity arised because, while our pipeline for executing visual lexicons (Section 6.2) is feasible for simpler cases, its limitations become apparent in complex scenarios, such as handling multiple subjects or intricate relationships, leaving significant room for improvement. We showcase some participant-created story using Brickify, including planting a tree (Figure 15), sharing an apple (Figure 16), and hosting a music party (Figure 17), where they crafted three-scene narratives effectively. However, failure cases (Figure 18) reveal that the model often omits subjects when there are more than three and/or when subjects overlap significantly (e.g., being "inside"). Additionally, the model sometimes fails to match specified sizes and occasionally produces patchy images. In response, participants typically regenerated outputs with different seeds, which sometimes worked. If not, they reduced the complexity, such as removing some subjects or tweaking the layout, to gradually adapt themselves to the model's capacity limits. We anticipate that as the base model (currently Stable Diffusion 2.1) continues to grow in size and evolve in architecture, its performance in executing the visual lexicon would improve, thus mitigating this problem.

8.7 Observations and Participants' Feedback

8.7.1 Decomposing a reference image into elements is more effective than using it as a whole. Participants consistently valued decomposing reference images into individual elements rather than using them as a whole, aligning with their design process. P3 noted, *"Being able to break down an image into parts lets me mix and match elements in a way that fits my vision, rather than being constrained by the original composition"*. Compared to their previous experiences with tools that only allow for remixing entire images, participants found the ability to recombine multiple decomposed elements and specify their relationships particularly valuable. P7 expressed, *"With other tools, I'd have to try different images multiple times to get something close to what I want. Here, I can just pull out the pieces I need and arrange them how I like"*.

8.7.2 Brickify enhances the sense of control but requires a target in mind. Participants unanimously agreed that Brickify offers

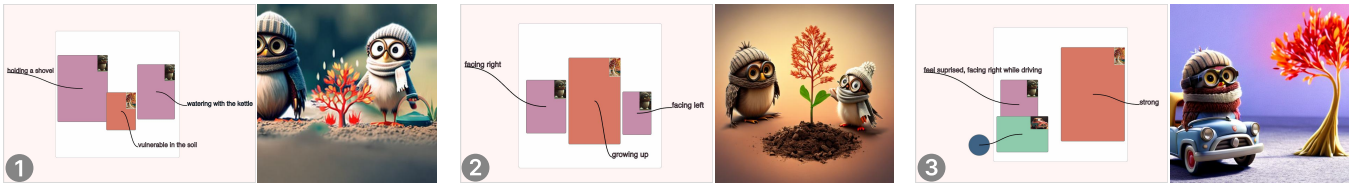

Figure 15: P3 generated results with Brickify, illustrating a story where (1) two owls plant a tree together, (2) nurture it with care, and (3) later, one owl drives by and happily witnesses the tree's growth into tall and strong.

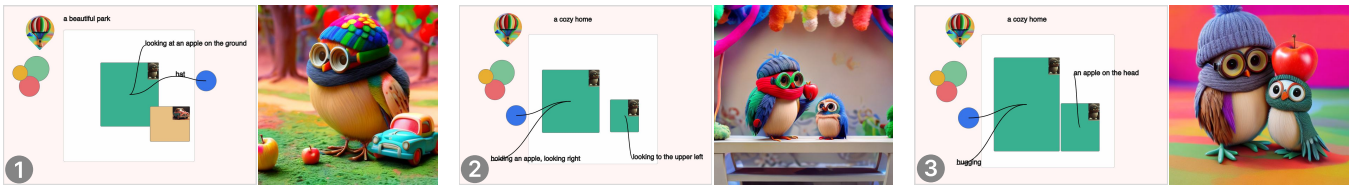

Figure 16: P6 generated results with Brickify. The user intends to describe (1) an owl discovers some apples in a park, picks one, and (2) brings it home to share with a friend, and (3) his friend puts the apple on the head, sharing a happy moment together.

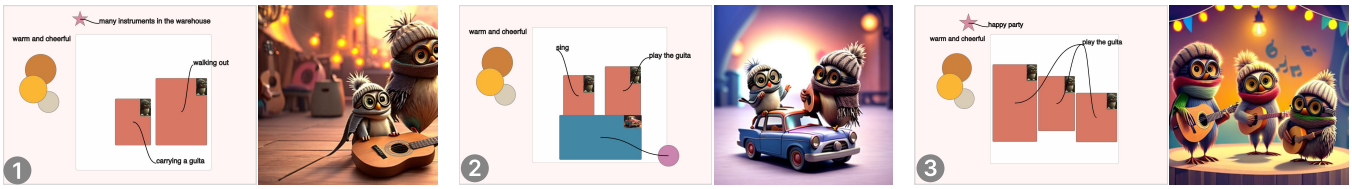

Figure 17: P10 generated results with Brickify, depicting a story where (1) the owl and his friend discover and move a guitar from a warehouse, (2) joyfully sing and play the guitar on a car, and (3) invite another friend to join them for a lively music party.

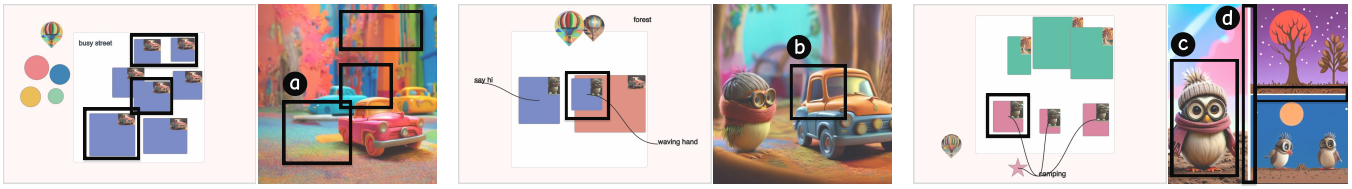

Figure 18: Failure cases: (a) when there are multiple subjects, the model often omits some of them (from P1); (b) when two subjects overlap significantly, the model struggles and incorrectly interprets the interaction as being "inside" (from P5); (c) the size of generated subjects sometimes may not align with the user-specified one; and (d) the model occasionally fails to produce a cohesive image, instead generating patchy outputs with visible edges (from P11).

a greater sense of control compared to using natural language prompts. They found it particularly useful for defining and manipulating the relationships between elements. P4 noted, *"it (Brickify) makes it so much easier to specify how different parts of my design interact. I feel like I have more direct control over the outcome"*. However, participants also highlighted that this increased control comes with a prerequisite: having a clear idea or target in mind. When they were unsure of what they wanted to create, they found that natural language prompts offered a quicker and more flexible starting point. As P9 explained, *"If I don't have a clear idea, it's easier to just throw in some random words and see what the AI generates, it's a good way to get inspired"*. These feedback implies a trade-off between control and exploration. Natural language allows for broad exploration and can spark new ideas even from vague or random inputs, while Brickify excels in depth and precision when users have a rough direction or specific visual properties in mind. As P1 put it, *"Once I know the general look I'm going for, it (Brickify) lets me really arrive there"*.

8.7.3 Direct manipulation on design tokens helps reduce the efforts of refinements. Brickify allows for more precise and enjoyable fine-tuning. Participants felt useful to directly manipulate elements without losing the core identity of their design. *"[With Brickify,] I can make the adjustments I want without compromising the overall look, I actually very enjoy this refinement process"*, shared by P5. The persistence of design tokens throughout the creative process was particularly valued. Participants liked that once a design token was created, it could be reused throughout a session, streamlining the workflow and ensuring consistency across different iterations.

"I love that I can just drag and drop a copy of a token for reuse," P8 shared, *"It keeps my design consistent without having to start from scratch every time"*. In contrast, participants mentioned the challenges they faced with other tools when trying to make minor refinements. In many cases, they found that the visual identity of their work would start to drift with each new text-based prompt, leading to frustration. *"I've tried making refinements in other tools, but often the visuals change too much, even when I just want minor tweaks. It's so frustrating that I usually just switch to Photoshop to fix it myself"*, P12 explained.

8.7.4 Brickify do not require a rigid structure thus enable a flexible thinking flow. Several (P1, P2, P4, P7, P8, P9, P11) participants mentioned they benefited from freely constructing their visual lexicon without worrying about ordering and format constraints. They felt Brickify can enable them to think non-linearly and creatively. P7 noted, *"with text prompts, I was advised to follow a template for effective prompting, though I don't know if it really matters: start with the overall context, list subjects, describe the view with terms like headshots or close-ups, and then add magic words for styling. It feels like I'm being forced to think in a certain way, which is not how design should work, it is supposed to be messy. You think about this, then that, then come back to adjust one accordingly, the elements are influencing each other"*. P11 expressed frustrations that *"different tools often require different prompt formats, making it (prompting with texts) more complicated"*. In contrast, Brickify does not assume a procedural process to construct the visual lexicon, users are free to start from any design aspects and seamlessly navigate between them. Overall, the flexibility offered by Brickify was seen as a significant advantage, enabling users to engage in a more dynamic and less restrictive creative thinking process.

8.7.5 Clarity in intent expression help users appropriately understand AI accountability. Users often face challenges in correctly attributing AI failures due to the lack of transparency in how AI interprets their inputs. This can lead to confusion, where users mistakenly believe that failures are due to unclear intent rather than AI's limitations. As P5 mentioned, *"Sometimes, I'm not sure if the problem is with how I'm phrasing things"*. This uncertainty often results in users repeatedly refining their inputs, leading to unnecessary back-and-forth. However, with Brickify, users felt more confident in the clarity of their intent expression, such confidence making users can appropriately attribute failures to AI. P12 commented that *"[With Brickify,] I know if something goes wrong, it's probably the AI, not me"*.

9 DISCUSSION

We reflect on the design of Brickify and discuss the lessons we learned and the implications for future research.

9.1 Design Implications

9.1.1 Integrating texts into visual-centric paradigm versus embedding visuals into texutal-centric paradigm. Brickify addresses two key ambiguities in textual prompts: *what* elements to use and *how* to construct them for an intended effect. Prior work, like DirectGPT [61], also augments textual prompts by embedding visual symbols to clarify object references, focusing on the *what* aspect. However, these two approaches differ fundamentally in their expressiveness of the *how* aspect due to their structural nature. Visual-centric paradigms like Brickify operate in a two-dimensional (2D) space, enabling spatial manipulation and richer exploration of relationships between elements, with text serving as a complement. In contrast, text-centric paradigms with embedded visuals, such as DirectGPT, function within a linear, one-dimensional (1D) space, where the narrative sequence is preserved, and visuals enhance object references or enrich textual information. We argue that the effectiveness of either paradigm is likely task-dependent. For visual tasks, the spatial properties of a 2D approach can provide more intuitive and efficient interactions, aiding in design, spatial reasoning, and layout organization. However, this may disrupt the narrative flow, making it less suitable for tasks that require step-by-step instructions such as programming. Conversely, a text-centric approach with visual enhancements may be better suited for tasks requiring logical reasoning and narrative coherence. This distinction raises important questions for future research: *when* should one paradigm be chosen over the other? Or is there potential for blending the spatial advantages of visuals with the narrative flow of text to enable a unified paradigm for various tasks?

9.1.2 Dynamic distributed agency between user and AI: letting users to specify when, where, and how much. Brickify is designed to reduce the ambiguity in design intent expression and thus improve users' sense of control when working with AI. We recognize that users' required control varied between individuals with different skill levels and varied at different stages in design process. When users seek to leverage AI's creativity, they often *choose* to leave ambiguity in the visual lexicon for the AI to refine. Conversely, when they have a clear vision, they specify their intentions with detailed visual tokens and intricate spatial manipulations, letting AI to execute their vision with precision. Thus, the agency distribution between user and AI is dynamically changing, also discussed by Satyanarayan et al. [75]. In Brickify, users can actively and explicitly configure *when*, *where*, and *how much* control they wish to shift to AI through the use of the imagination token. For instance, during the Study 2, P2 provided a brief description "a beautiful park" as background, and assign a large imagination token, signaling the AI to elaborate. Participants in our study valued this flexibility and control. In contrast, most current text-to-image tools, like DaLLE, MidJourney, and Adobe Firefly, automatically expand and refine users' whole prompts without asking users if they want, leading to unintended results. This informs the importance of providing an *explicit* way for users to actively delegate control to AI — managing *when*, *where*, and *how much* — rather than assuming a fixed agency distribution pre-defined by the system.

9.1.3 Towards bi-directional visual lexicon construction: enabling both users and AI to be constructors. Reflecting the current design of Brickify, design tokens serve as the communication medium between users and AI, while the constructed visual lexicon acts as a visual abstraction of the generated image. However, Brickify only allows users to construct the visual lexicon, with AI solely acting as the receiver to execute it. This workflow presents a challenge: as mentioned by some participants, in the early ideation stage when users may not have a clear vision, they would prefer natural language as a quick starting point. But how can users complete the

iterative design process without an initial visual lexicon? What if they wish to refine an AI-generated image instead? The current system of Brickify does not fully close this interaction loop. One potential solution is to enable AI to construct the visual lexicon as well. Given an image, AI could automatically extract design tokens and compose a corresponding visual lexicon for users to manipulate. This approach is similar to the prior work on *abstraction-driven* color manipulation for image [76] and motion graphic videos [77], where the system creates color abstractions for manipulation. This informs that visual abstractions can be constructed bi-directionally, with both users and AI acting as constructors. However, this also raises the questions warrant more in-depth exploration in future work: how to decide the granularity of the AI-generated visual lexicon; what elements should be reified into tokens; and what relationships should be reflected?

9.2 Design Opportunities in Brickify

9.2.1 Diversifying design token types and sources. Currently, Brickify supports limited visual token types including subject, style, and color. However, there are more visual elements essential for constructing a successful design, such as camera angle, depth, texture, and material. Brickify can naturally incorporate them by adding corresponding token types. For example, with a camera token, users can specify the camera angle by positioning it in the visual lexicon, which is otherwise hard to express with texts. It is also possible to automate the token creation process by decomposing an image into low-level design elements [91]. It is worth noting that Brickify currently only supports reference image as the source of visual tokens. While starting from references is common, it is certainly not the only approach that designers use. Designers could draw inspirations from other sources such as their own memory, using sketches to externalize and articulate their intent [46, 50]. Expanding Brickify to incorporate sketch as a source of design tokens is a promising way to accommodate such scenarios.

9.2.2 Customizing and Re-configuring design tokens by users. While Brickify can expand the supported design token types, it is impractical to preset every possible type. A valuable future direction is to allow users to make their own tokens. Users could define tokens by describing their functions in natural language and providing examples. The system would then dynamically generate and support these tokens. Furthermore, preset tokens may not always match user's intended usage. It would be more flexible to allow users to appropriate the pre-designed tokens for their specific needs — reconfiguring their represented meanings. For example, while current color tokens only convey proportional information through size, we observed P3 using their position to indicate specific areas on a subject to colorize. Lastly, it is important to constrain such customization and reconfiguration to remain interpretable by generative AI models to ensure effective interaction.

9.2.3 Affording richer manipulations on design tokens. Brickify currently supports the manipulations of resizing, positioning, grouping, and linking. However, during the user study, participants attempted additional manipulations beyond those provided. For example, in the first task, many participants (7/12) tried to rotate the token to indicate the pumpkin's position. Similarly, some (3/12) wanted to move subject tokens forward or backward to specify spatial relationships. These observations reveal the potential for our interaction paradigm to support more intricate user intentions. Beyond spatial relationships, other behaviors like blending could also be supported. While richer manipulations can enhance user experience, they also introduce complexity. Ideally, users should rely on *technical reasoning* — intuitively understanding how manipulations affect outcomes — rather than *procedure learning* — memorizing steps to achieve desired effects [9]. Future work will investigate what manipulations are both desired and natural for users to express their intentions effectively.

9.2.4 Propagating the modifications of the design token. A limitation of the current version of Brickify is that only textual tokens are editable, while visual tokens, once created, cannot be re-linked to another visual element. For example, if a user creates a subject token of an owl and uses it in multiple designs, changing this owl to a rabbit requires reconstructing everything from scratch. To address this, a possible improvement is to allow design token modification with automatic propagation to all instances. Since we designed the visual tokens to be persistent and every time users construct a visual lexicon, they drag a copy of the original token, this creates a natural link between the original and its copies. Leveraging this link, once users modify the original token, such as re-attaching it to a new subject in another image, we can propagate this modification to all its copies in related visual lexicons. Such a propagation mechanism can allow designers to quickly compare different visual candidates and streamline their workflow.

9.2.5 Beyond static graphic design: extending Brickify to video, 3D scene, and other co-creation tasks with AI. While Brickify is initially designed for static graphic design, its visual-centric interaction paradigm holds significant potential for broader applications in general AI-assisted visual design tasks. In video creation, for instance, Brickify could adapt the visual lexicon to a timeline-based structure, where individual lexicons construct each scene. Similarly, in 3D scene modeling, where spatial relationships are more complex, Brickify could extend 2D design tokens into 3D tokens and extend the 2D manipulations to 3D operations. While token design and manipulations may be domain-specific, the fundamental interaction logic of using direct manipulation on tokens to construct elemental relationships remains coherent and consistent across different design domains. We believe Brickify opens up the possibility to offer designers a unifying design language to communicate with AI across the broad creative landscape.

9.3 Limitations

9.3.1 Brickify might fail in describing unseen visuals beyond recombination. Each modality has special strengths and weaknesses in its ability to communicate particular concepts [12]. While Brickify excels at referring to elements and describing spatial relationships, its reliance on existing visuals might lead to design fixation [46], where designers may unconsciously adhere to what has already been known or available. In contrast, natural language could describe unseen visuals that go beyond recombination (e.g., *"a cute sock with a human-like face"*) or ideas that might seem unreasonable (e.g., *"time is melting"*). Therefore, our proposed visual-centric

interaction paradigm is not intended to replace the text-centric approach. Instead, it is important for designers to strategically choose the most suitable modality based on the design context and the level of originality they seek to achieve.

9.3.2 Visual lexicon extraction could be improved. Our implementation of visual lexicon execution relies heavily on off-the-shelf computer vision techniques. Despite that we selected the state-of-the-art ones at the time of developing Brickify, these techniques have inherent limitations that impact our system's capacity. Currently, Brickify supports only 4-6 subject tokens. This constraint arises because we use the `Break-A-Scene` [5] approach to preserve each subject's visual identity. However, our experiments show that when there are more than six subjects, the performance in maintaining visual identity drops significantly. The scope of this work is not to improve the performances of computer vision models, but we do hope this work informs the importance of computer vision research to push the progress forward — training the model to be aware and preserve visual details rather than solely taking natural language as input.

9.3.3 Inference and computation costs could hinder user experience. Due to the high computational costs and inference time of diffusion models, we cannot support on-the-fly inference and immediate feedback. Our visual lexicon execution pipeline sequentially handles different design aspects (layout, style, and color), each requiring a 50-step diffusion model inference (around 30 seconds). As a result, users must click a button to generate and wait for the results, interrupting their design experience to some degree. We envision that as inference time and computation costs decrease, users will no longer need to click the generation button after constructing the visual lexicon. Instead, they will receive immediate feedback while manipulating the tokens. This will enable users to instantly see the effects of their actions, providing a smoother co-creation experience with generative models.

9.3.4 Study results might not be generalizable for design novices. Participants in our study are experienced designers, all with a minimum of 3 years of design experience. These participants are trained to approach design problems visually, and Brickify was specifically designed to align with this visual-centric mental model. As a result, the study's findings may not generalize to novice designers or casual users who lack this level of expertise. It is uncertain whether novice users could adapt to this visual-centric paradigm and fully leverage the fine-grained control.

10 CONCLUSION

In this paper, we introduce Brickify, a *visual-centric* interaction paradigm that allows users to express design intent more effectively. Brickify reifies primitive design elements from reference images into *interactive*, *reusable* design tokens, enabling users to specify *what* elements to use and *how* to construct them towards the desired effect. We implement Brickify to exemplify how state-of-the-art AI models can execute users' intent expressed through Brickify. In a user study, experienced designers found it easier to describe visual details and relationships with fewer mental demands through Brickify. They efficiently explored design alternatives by reusing tokens and performed refinements more quickly, particularly for complex designs. Designers preferred Brickify over textual-centric prompting approach, valuing the sense of control it provided when they had design ideas in mind. Moving forward, we plan to extend Brickify to include more element types and operations, broadening its expressive capabilities. The design implications derived from this work shed light on future research to design effective interaction mediums for human-AI co-creation.

ACKNOWLEDGMENTS

We sincerely thank Xueguang Ma, Ryan Yen, and anonymous reviewers for their insightful suggestions that have led to a great improvement of this work. We also extend our appreciation to Annie Sun and Hakeerat Singh Mayall for their kind help on a few early-stage prototypes. We would also like to thank all our participants for their time and valuable input. This work is supported in part by the Discovery Grant (RGPIN-2020-03966) from the NSERC (Natural Sciences and Engineering Research Council of Canada) and a gift fund from Adobe Systems Inc.

We acknowledge that much of our work takes place on the traditional territory of the Neutral, Anishinaabeg, and Haudenosaunee peoples. Our main campus is located on the Haldimand Tract, the land granted to the Six Nations that includes six miles on each side of the Grand River.

REFERENCES

[1] Mahmoud Afifi, Marcus A Brubaker, and Michael S Brown. 2021. Histogan: Controlling colors of gan-generated and real images via color histograms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7941–7950.

[2] Flux AI. 2024. https://flux-ai.io/ Accessed: November 22, 2024.

[3] Tyler Angert, Miroslav Suzara, Jenny Han, Christopher Pondoc, and Hariharan Subramonyam. 2023. Spellburst: A Node-based Interface for Exploratory Creative Coding with Natural Language Prompts. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–22.

[4] Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L Glassman. 2024. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[5] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. 2023. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*. 1–12.

[6] Omri Avrahami, Ohad Fried, and Dani Lischinski. 2023. Blended latent diffusion. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–11.

[7] Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18208–18218.

[8] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. 2022. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*. Springer, 707–723.

[9] 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.

[10] 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*. 102–109.

[11] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf* 2, 3 (2023), 8.

[12] Bill Buxton. 1989. The "Natural" language of interaction: A perspective on non-verbal dialogues. *INFOR: Information Systems and Operational Research* 27, 2 (1989), 221–229.

[13] Xiang'Anthony' Chen, Jeff Burke, Ruofei Du, Matthew K Hong, Jennifer Jacobs, Philippe Laban, Dingzeyu Li, Nanyun Peng, Karl DD Willis, Chien-Sheng Wu, et al. 2023. Next steps for human-centered generative ai: A technical perspective. *arXiv preprint arXiv:2306.15774* (2023).

[14] Erin Cherry and Celine Latulipe. 2014. Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on*

Computer-Human Interaction (TOCHI) 21, 4 (2014), 1–25.
[15] Lydia B Chilton, Ecenaz Jen Ozmen, Sam H Ross, and Vivian Liu. 2021. VisiFit: Structuring iterative improvement for novice designers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.
[16] Lydia B Chilton, Savvas Petridis, and Maneesh Agrawala. 2019. Visiblends: A flexible workflow for visual blends. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
[17] 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*. 1–25.
[18] John Joon Young Chung and Eytan Adar. 2023. Artinter: AI-powered Boundary Objects for Commissioning Visual Arts. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 1997–2018.
[19] 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*. 1–17.
[20] 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*. 1–19.
[21] Niv Cohen, Rinon Gal, Eli A Meirom, Gal Chechik, and Yuval Atzmon. 2022. "This is my unicorn, Fluffy": Personalizing frozen vision-language representations. In *European conference on computer vision*. Springer, 558–577.
[22] ComfyUI. 2024. https://www.comfy.org/en/ Accessed: November 22, 2024.
[23] Nigel Cross. 1990. The nature and nurture of design ability. *Design studies* 11, 3 (1990), 127–140.
[24] OpenAI DALL·E3. 2024. https://openai.com/index/dall-e-3/ Accessed: November 22, 2024.
[25] Donis A Dondis. 1974. *A primer of visual literacy*. Mit Press.
[26] Claudia Eckert and Martin Stacey. 2000. Sources of inspiration: a language of design. *Design studies* 21, 5 (2000), 523–538.
[27] Lisa A Elkin, Matthew Kay, James J Higgins, and Jacob O Wobbrock. 2021. An aligned rank transform procedure for multifactor contrast tests. In *The 34th annual ACM symposium on user interface software and technology*. 754–768.
[28] Nada Endrissat, Gazi Islam, and Claus Noppeney. 2016. Visual organizing: Balancing coordination and creative freedom via mood boards. *Journal of Business Research* 69, 7 (2016), 2353–2362.
[29] John M Findlay and Iain D Gilchrist. 2003. *Active vision: The psychology of looking and seeing*. Number 37. Oxford University Press.
[30] Ronald A Finke, Thomas B Ward, and Steven M Smith. 1996. *Creative cognition: Theory, research, and applications*. MIT press.
[31] Adobe Firefly. 2024. https://www.adobe.com/ca/products/firefly.html Accessed: November 22, 2024.
[32] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations*.
[33] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–13.
[34] Steve Garner and Deana McDonagh-Philp. 2001. Problem interpretation and resolution via visual stimuli: the use of 'mood boards' in design education. *Journal of Art & Design Education* 20, 1 (2001), 57–64.
[35] Frederic Gmeiner, Humphrey Yang, Lining Yao, Kenneth Holstein, and Nikolas Martelaro. 2023. Exploring challenges and opportunities to support designers in learning to co-create with AI-based manufacturing design tools. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.
[36] Gabriela Goldschmidt and Dan Tatsa. 2005. How good are good ideas? Correlates of design creativity. *Design studies* 26, 6 (2005), 593–611.
[37] Charles Goodwin. 2015. Professional vision. In *Aufmerksamkeit: Geschichte-theorie-empirie*. Springer, 387–425.
[38] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. 2023. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7323–7334.
[39] Eleanor R Heider. 1972. Universals in color naming and memory. *Journal of experimental psychology* 93, 1 (1972), 10.
[40] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. 2022. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *The Eleventh International Conference on Learning Representations*.
[41] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. 2024. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4775–4785.
[42] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
[43] Josh Holinaty, Alec Jacobson, and Fanny Chevalier. 2021. Supporting reference imagery for digital drawing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2434–2442.
[44] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. 2023. Composer: creative and controllable image synthesis with composable conditions. In *Proceedings of the 40th International Conference on Machine Learning*. 13753–13773.
[45] Alexander Ivanov, David Ledo, Tovi Grossman, George Fitzmaurice, and Fraser Anderson. 2022. MoodCubes: Immersive spaces for collecting, discovering and envisioning inspiration materials. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*. 189–203.
[46] Ben Jonson. 2005. Design ideation: the conceptual sketch in the digital age. *Design studies* 26, 6 (2005), 613–624.
[47] Youwen Kang, Zhida Sun, Sitong Wang, Zeyu Huang, Ziming Wu, and Xiaojuan Ma. 2021. MetaMap: Supporting visual metaphor ideation through multi-dimensional example-based exploration. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
[48] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6007–6017.
[49] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.
[50] David Kirsh. 2010. Thinking with external representations. *AI & society* 25 (2010), 441–454.
[51] Janin Koch, Andrés Lucero, Lena Hegemann, and Antti Oulasvirta. 2019. May AI? Design ideation with cooperative contextual bandits. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
[52] Janin Koch, Nicolas Taffin, Michel Beaudouin-Lafon, Markku Laine, Andrés Lucero, and Wendy E Mackay. 2020. Imagesense: An intelligent collaborative ideation tool to support diverse human-computer partnerships. *Proceedings of the ACM on human-computer interaction* 4, CSCW1 (2020), 1–27.
[53] Janin Koch, Nicolas Taffin, Andrés Lucero, and Wendy E Mackay. 2020. SemanticCollage: enriching digital mood board design with semantic labels. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. Association for Computing Machinery, New York, NY, USA, 407–418.
[54] Krea.ai. 2024. https://www.krea.ai/home Accessed: November 22, 2024.
[55] Chinmay Kulkarni, Stefania Druga, Minsuk Chang, Alex Fiannaca, Carrie Cai, and Michael Terry. 2023. A word is worth a thousand pictures: Prompts as ai design material. *arXiv preprint arXiv:2303.12647* (2023).
[56] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-Concept Customization of Text-to-Image Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
[57] Vivian Liu, Jo Vermeulen, George Fitzmaurice, and Justin Matejka. 2023. 3DALL-E: Integrating text-to-image AI in 3D design workflows. In *Proceedings of the 2023 ACM designing interactive systems conference*. 1955–1977.
[58] Andrés Lucero. 2012. Framing, aligning, paradoxing, abstracting, and directing: how design mood boards work. In *Proceedings of the Designing Interactive Systems Conference* (Newcastle Upon Tyne, United Kingdom) *(DIS '12)*. Association for Computing Machinery, New York, NY, USA, 438–447. https://doi.org/10.1145/2317956.2318021
[59] Andrés Lucero and JBOS Martens. 2005. Mood Boards: Industrial designers' perception of using mixed reality. In *Proc. SIGCHI. NL Conference*. 13–16.
[60] 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*. 1–13.
[61] 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*. 1–16.
[62] MidJourney. 2024. https://www.midjourney.com/home Accessed: November 22, 2024.
[63] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4296–4304.
[64] Gregory Murphy. 2004. *The big book of concepts*. MIT press.
[65] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*. PMLR, 16784–16804.
[66] François Osiurak, Christophe Jarry, Philippe Allain, Ghislaine Aubin, Frédérique Etcharry-Bouyx, Isabelle Richard, Isabelle Bernard, and Didier Le Gall. 2009. Unusual use of objects after unilateral brain damage. The technical reasoning

model. *Cortex* 45, 6 (2009), 769–783.
[67] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. 2023. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.
[68] 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*. 804–818.
[69] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.
[70] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*. Pmlr, 8821–8831.
[71] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
[72] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22500–22510.
[73] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* 35 (2022), 36479–36494.
[74] Vishnu Sarukkai, Lu Yuan, Mia Tang, Maneesh Agrawala, and Kayvon Fatahalian. 2024. Block and Detail: Scaffolding Sketch-to-Image Generation. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 33, 13 pages. https://doi.org/10.1145/3654777.3676444
[75] Arvind Satyanarayan and Graham M. Jones. 2024. Intelligence as Agency: Evaluating the Capacity of Generative AI to Empower or Constrain Human Action. *An MIT Exploration of Generative AI* (mar 27 2024). https://mit-genai.pubpub.org/pub/94y6e0f8.
[76] Xinyu Shi, Mingyu Liu, Ziqi Zhou, Ali Neshati, Ryan Rossi, and Jian Zhao. 2024. Exploring interactive color palettes for abstraction-driven exploratory image colorization. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.
[77] Xinyu Shi, Yinghou Wang, Yun Wang, and Jian Zhao. 2024. Piet: Facilitating Color Authoring for Motion Graphics Video. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
[78] Yang Shi, Tian Gao, Xiaohan Jiao, and Nan Cao. 2023. Understanding design collaboration between designers and artificial intelligence: a systematic literature review. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–35.
[79] Ben Shneiderman. 1983. Direct manipulation: A step beyond programming languages. *Computer* 16, 08 (1983), 57–69.
[80] Auste Simkute, Lev Tankelevitch, Viktor Kewenig, Ava Elizabeth Scott, Abigail Sellen, and Sean Rintel. 2024. Ironies of Generative AI: Understanding and Mitigating Productivity Loss in Human-AI Interaction. *International Journal of Human–Computer Interaction* (2024), 1–22.
[81] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. 2023. StyleDrop: Text-to-Image Generation in Any Style. In *37th Conference on Neural Information Processing Systems (NeurIPS)*. Neural Information Processing Systems Foundation.
[82] Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* 32 (2019).
[83] 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*. 1–19.
[84] Kaiber Superstudio. 2024. https://www.kaiber.ai/product/ Accessed: November 22, 2024.
[85] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The metacognitive demands and opportunities of generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–24.
[86] Anne Tomes, Caroline Oates, and Peter Armstrong. 1998. Talking design: negotiating the verbal–visual translation. *Design Studies* 19, 2 (1998), 127–142.
[87] Katja Tschimmel. 2012. Design Thinking as an effective Toolkit for Innovation. In *ISPIM Conference Proceedings*. The International Society for Professional Innovation Management (ISPIM), 1.
[88] Tiffany Tseng, Ruijia Cheng, and Jeffrey Nichols. 2024. Keyframer: Empowering Animation Design using Large Language Models. https://arxiv.org/abs/2402.06071
[89] Priyan Vaithilingam, Elena L Glassman, Jeevana Priya Inala, and Chenglong Wang. 2024. DynaVis: Dynamically Synthesized UI Widgets for Visualization Editing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
[90] Veera Vimpari, Annakaisa Kultima, Perttu Hämäläinen, and Christian Guckelsberger. 2023. "An Adapt-or-Die Type of Situation": Perception, Adoption, and Use of Text-to-Image-Generation AI by Game Industry Professionals. *Proceedings of the ACM on Human-Computer Interaction* 7, CHI PLAY (2023), 131–164.
[91] Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. 2023. Concept decomposition for visual exploration and inspiration. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–13.
[92] Qian Wan and Zhicong Lu. 2023. GANCollage: A GAN-Driven Digital Mood Board to Facilitate Ideation in Creativity Support. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 136–146.
[93] Sitong Wang, Savvas Petridis, Taeahn Kwon, Xiaojuan Ma, and Lydia B Chilton. 2023. PopBlends: Strategies for conceptual blending with large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.
[94] 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*. 1–21.
[95] Anthony Ward. 1984. Design cosmologies and brain research. *Design Studies* 5, 4 (1984), 229–238.
[96] Colin Ware. 2010. *Visual thinking for design*. Elsevier.
[97] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15943–15953.
[98] Justin D Weisz, Jessica He, Michael Muller, Gabriela Hoefer, Rachel Miles, and Werner Geyer. 2024. Design Principles for Generative AI Applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–22.
[99] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 143–146.
[100] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–22.
[101] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. 2023. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7452–7461.
[102] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.
[103] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. 2017. Real-time user-guided image colorization with learned deep priors. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–11.

Table 1: The table records participants' demographic information, including gender, age, occupation, and experiences of graphic design in years (Design Exp.), self-rated frequency of Generative AI usage (GenAI Freq.), and GenAI tools they frequently use.

ID	Gender	Age	Occupation	Design Exp.	GenAI Freq.	GenAI Tools
P1	F	25	Design New Grad	5	3	Midjourney
P2	M	27	Design Researcher	5	5	Dall-E
P3	F	Not Reveal	Designer	4	3	Midjourney
P4	M	28	Visual Designer	5	3	Midjourney
P5	F	25	3D Artist	5	5	Dall-E, Midjourney
P6	F	26	Exhibition Designer	4	3	MidJourney
P7	M	27	Graphic Designer	5	2	MidJourney
P8	F	26	UX Designer	5	4	MidJourney
P9	M	25	Technical Aritist	3	3	MidJourney
P10	M	25	Architect	3	2	Dall-E
P11	M	25	Graphic Designer	5	4	MidJourney
P12	M	30	Visual Designer	5	3	MidJourney

Table 2: Rating rubric for the quality of participants' expressions in terms of the element coverage, clarity of size, position, style, and color. Raters were rated on a 7-point Likert Scale where 1 means very low and 7 means very high.

Item	Score	Criteria
Element Coverage	7	All key elements from the target image are included and accurately represented.
	4	Three elements are missing.
	1	Six or more elements are missing.
Clarity of Size	7	The relative size of all elements is clearly and accurately described.
	4	The relative size of around half elements is described, but some ambiguity exists.
	1	The size of elements is highly unclear or not described.
Clarity of Position	7	The position of all elements is clearly and accurately described relative to each other.
	4	The position of around half elements is described, but some spatial relationships are ambiguous.
	1	The position of elements is highly unclear or not described.
Clarity of Style	7	The global style is clearly and accurately described. 1) Referred to the style of the first and/or the second reference images; or 2) style descriptors such as "minimal/simplicity/geometric/abstract" or similar ones.
	4	The global style is described, but it is not apparent how it relates to the target image.
	1	The style of elements is unclear or not described.
Clarity of Color	7	The color of all elements is clearly and accurately described.
	4	The color of about half elements is described and close to the target colors, but the rest are ambiguous or missing.
	1	The color of elements mostly described far away from the target image or not described.

A APPENDIX

A.1 Implementation Details

The front-end user interface of Brickify was built using React.js as the primary framework. Most of the design token management functionalities, such as the creation, deletion, and manipulation, were implemented using D3.js. The rest of the interface components, such as the buttons and icons, were taken from the Material UI library and customized to fit the needs of the application. The server-side rendering for API calls is handled by fastAPI. The back-end model fine-tuning and inferences for visual lexicon execution are written in Python and performed on an 80G A100 GPU.

A.2 Participants' Demographic Information

Table 1 describes the detailed demographic information of participants in our user study.

A.3 Rating Rubric for Expressions in Study 1

Table 2 lists the detailed rating rubric for the external scorers in Study 1 to rate participants' expressed intention.