GenFODrawing: Supporting Creative Found Object Drawing with Generative AI

Jiaye Leng, Hui Ye, Pengfei Xu, Miu-Ling Lam, and Hongbo Fu

Abstract—Found object drawing is a creative art form incorporating everyday objects into imaginative images, offering a refreshing and unique way to express ideas. However, for many people, creating this type of work can be challenging due to difficulties in generating creative ideas and finding suitable reference images to help translate their ideas onto paper. Based on the findings of a formative study, we propose *GenFODrawing*, a creativity support tool to help users create diverse found object drawings. Our system provides AI-driven textual and visual inspirations, and enhances controllability through sketch-based and box-conditioned image generation, enabling users to create personalized outputs. We conducted a user study with twelve participants to compare *GenFODrawing* to a baseline condition where the participants completed the creative tasks using their own desired approaches without access to our system. The study demonstrated that *GenFODrawing* enabled easier exploration of diverse ideas, greater agency and control through the creative process, and higher creativity support compared to the baseline. A further open-ended study demonstrated the system's usability and expressiveness, and all participants found the creative process engaging.

Index Terms—Creativity support tool, idea generation, creative drawing, interactive creation, generative AI.

1 Introduction

Found objects, everyday items discovered by chance and imbued with special meaning, challenge conventional perceptions and serve as unique mediums for emotional expression [1]. This creative manner of repurposing objects enhances mood, memory, cognition [2], and shows therapeutic benefits [3]. Found object drawing, as a creative process of incorporating everyday objects into visual compositions, enables people to integrate items from their surroundings, such as household items, foods, and shadows, as part of their drawings (Figure 1). This creates an interesting interaction between the object's physical form and drawn elements, blurring the boundary between reality and imagination. The combination of the found objects and the drawn lines encourages people to transform ordinary items into elements of imaginative expression. Beyond its artistic value, found object drawing serves as an effective approach that enhances creative thinking and cognitive development, finding applications in various domains such as education [4] and therapeutic practices [5].

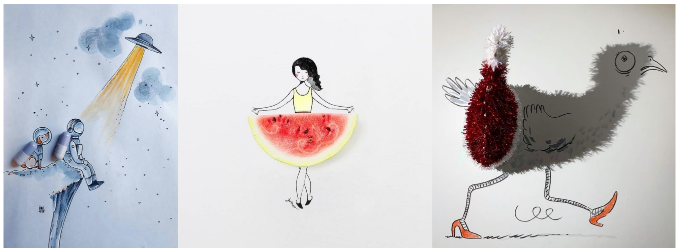


Fig. 1. Found object drawings from Pinterest: they are drawn with capsules, a watermelon slice, and a shadow, creating object-level or scene-level compositions.

Currently, creating found object drawings relies heavily on individual imagination and manual exploration, often requiring iterative attempts to discover meaningful object-drawing combinations. This trial-and-error process can be time-consuming and cognitively demanding, particularly for those new to this art form or struggling with creative ideation. Existing creativity support tools primarily focus on either facilitating the drawing process and results [6]–[11], or generating visual inspirations through recombining and transforming existing elements [12]–[14]. However, found object drawing presents unique challenges that current tools do not adequately address. Unlike traditional drawing or design tasks, it requires users to both identify creative associations for everyday objects and skillfully incorporate these physical objects into works. It means that users need to imaginatively interpret found objects as something different from themselves while maintaining visual plausibility. This creative process demands support for associative thinking [15], [16] and practical guidance in object integration, which remains unexplored by existing tools.

In order to further investigate users' needs in creating found object drawings, we conducted a formative study with six participants. The findings revealed several key insights. First, the participants identified associative thinking, i.e., connecting found objects with other objects, as a starting point of the creative process. Second, the analysis of their drawing results showed a division between object-level and scene-level compositions, with varying preferences among

Corresponding author: Hongbo Fu.
Jiaye Leng and Miu-Ling Lam are with the School of Creative Media, City University of Hong Kong. E-mail: jiayeleng2-c@my.cityu.edu.hk, miu.lam@cityu.edu.hk.
Hui Ye is with the Department of Interactive Media, School of Communication, Hong Kong Baptist University, and the Division of Arts and Machine Creativity, Hong Kong University of Science and Technology. E-mail: huiyehy@hkbu.edu.hk.
Pengfei Xu is with the College of Computer Science and Software Engineering, Shenzhen University. E-mail: xupengfei.cg@gmail.com.
Hongbo Fu is with the Division of Arts and Machine Creativity, Hong Kong University of Science and Technology. E-mail: hongbofu@ust.hk.

the participants. Lastly, despite differences in drawing skills, the participants encountered similar challenges in translating their creative ideas onto paper. They expressed a need for reference images that are aligned with their creative intentions, which are often difficult to obtain.

Based on the findings from our formative study, we introduce *GenFODrawing*, a creativity support tool that leverages multiple generative models, including Visual Language Model (VLM), Large Language Model (LLM), Text-to-Image (T2I) model, and Image-to-Image (I2I) model. This tool aids users in creating found object drawings with both textual and visual inspirations. *GenFODrawing* helps users discover creative possibilities by analyzing the input object and generating diverse associations within user-specified themes and drawing area constraints. Users can explore various visual compositions at both object and scene levels, offering visual descriptions enriched with moods. The system allows for customization of reference images through controllable text-to-image generation, guided by user style preferences, sketch inputs, and bounding boxes, which enable users to precisely control the placement and integration of the found object within the generated image. All generated references can be further converted into drawing-friendly formats, supporting the transition from digital ideation to manual drawing. Through this interactive, human-in-the-loop workflow, *GenFODrawing* provides comprehensive support for found object drawing, from initial ideation to final execution.

We compared *GenFODrawing* with a baseline condition in which participants completed creative tasks using their own desired approaches and without access to our system. The results demonstrated that participants reported significantly higher scores in creativity support and controllability for *GenFODrawing* than the baseline. A subsequent open-ended study showed that all participants successfully created a variety of found object drawings, and reported high usability and low workload when using our system. Several participants emphasized the importance of maintaining user autonomy and active participation, rather than passively accepting AI outputs. Thus, effective creativity support tools should balance the benefits of AI assistance with mechanisms that foster and preserve human creative engagement [17], [18].

In summary, the contributions of our work are threefold:

- A formative study that identifies the challenges and needs of users when creating found object drawings.
- *GenFODrawing*: a novel creativity support tool that integrates multiple LLMs and introduces sketch-based and box-conditioned image generation for controllable and user-aligned outputs. It facilitates associative thinking with found objects, provides creative and personalized reference images, and helps users seamlessly incorporate found objects into their drawings.
- Two user studies: a comparative study with a baseline system involving 12 participants, demonstrating the effectiveness of *GenFODrawing*, and an open-ended study validating the tool's usability.

2 Related Work

2.1 Creativity Support Tool (CST)

The creative process typically involves alternating divergent and convergent thinking to generate new ideas [19]–[21]. To support this process, researchers have developed CSTs [22], which integrate computational and interactive technologies to facilitate creative processes and prevent fixation [23]. Some efforts have focused on implementing drawing, exploring how to facilitate people's drawing experience [6]–[8] or how to get better drawing results [9], [10], while paying less attention to the creative ideas before drawing. Recent works have attempted to assist users in idea generation through generated images. These approaches leverage generative models to provide visual inspirations, focusing on style transfer [12], [13], semantic variations [24], or recombination of creative elements [14]. However, found object drawing poses unique challenges unexplored by current CSTs: it requires both breaking conventional thinking patterns to associate semantically unrelated objects and seamlessly integrating physical objects into creative compositions. Such requirements make it difficult to support with either existing visual search [25] for associative inspiration or standard image generation for drawing guidance.

A closely related system, MetaMap [26], facilitates visual metaphor ideation by enabling associative exploration using example images retrieved from large datasets based on semantic, color, or shape similarity. Our work differs from MetaMap in several key aspects: (1) we leverage recent VLMs to directly generate shape-based associations, allowing for suggestions unconstrained by dataset coverage; (2) our system emphasizes conceptual association rather than merely retrieving images, encouraging broader idea generation; and (3) associative thinking is only the starting point in our interactive pipeline, which further supports users through visual description generation and reference image synthesis with generative models. This multi-stage approach more fully involves users in the creative process. Beyond images, textual materials can also stimulate creative thinking through abstraction. This is exemplified in mood boards, which traditionally combine both visual and textual elements. Recent works [13], [27] leveraged AI techniques to enhance mood boards and support ideation. Inspired by these works, we developed the first AI-driven CST for creative found object drawing, leveraging both texts and images for idea generation.

2.2 Human-AI Co-Creation

The concept of co-creation has been explored for decades. Licklider et al. [28] proposed the concept of human-computer symbiosis, which means cooperative interaction between humans and computers, and they suggest that this symbiotic partnership would be significantly more effective than humans performing intellectual operations independently. With the rapid advancement of AI technology, co-creation between humans and AI has become a prominent topic. So far, existing works have explored human-AI co-creation topics across various domains, achieving notable progress, such as music [29], video [30], [31], writing [32], [33], programming tutorials [34], sports news [35], design [36], [37], etc. Some studies indicate that AI-driven co-creation has significant potential in enhancing creative ideation. Liao et al. [38] discussed that AI could provide inspiration, widen design scope, or trigger design actions by suggesting texts or images. Kim et al. [39] showed that AI's involvement in design ideation could effectively slow the decline in novelty, variety, and quantity of ideas. However, recent work suggests these gains may come at a cost. Kumar et al. [18] found that although LLM assistance yields short-term boosts in both divergent and convergent tasks, it can actually undermine people's independent creative performance when AI support is removed. Doshi et al. [40] also showed that generative AI may enhance individual creativity but drive homogenization and reduce collective novelty. Building on extensive human–AI co-creation research, our system empowers users with multiple control dimensions to overcome idea homogenization and assert agency in the creative process of found object drawing.

2.3 Intent Expression in Generative Models

Although LLMs can produce diverse outputs, generating results that align with user intentions remains challenging. To meet users' expectations, prompt engineering has become an important strategy. Chain-of-thought reasoning [41] is a widely used approach that guides models to analyze and solve problems step by step. Providing instructions and examples can also help the model understand users' expected outputs, known as in-context learning [42], [43]. In the case of T2I models, researchers have focused on controllable generation to achieve satisfactory visual outputs. Existing works attempted to enhance the controllability of results by introducing additional input conditions, including sketches [44], segmentation masks [45], bounding boxes [46], etc. Our work explores the integration of these generative techniques into the creative process of found object drawing, enabling VLMs and LLMs to help users explore ideas that align with their creative intentions, as well as using a sketch-based T2I model to generate precise reference images and an I2I model to convert them to pencil sketches for drawing guidance.

3 Formative Study

To investigate the challenges in creating found object drawings and identify user's needs for a creativity support tool, we conducted a formative study with an emphasis on seeking the answers to the following questions:

- How do users create found object drawings?
- What challenges do users encounter in the creative process of found object drawings?

After the study, we summarized important findings and formulated our system design goals.

3.1 Participants

Our recruitment aimed to include both novices and design professionals to understand how users with different expertise levels approach this creative task. We recruited six participants (4 males and 2 females, P1 to P6) aged between 24 and 30 (M = 26.7, SD = 2.73). They were PhD students from various fields, including computer science (P1), human-computer interaction (P2 and P3), communication studies (P4), and design (P5 and P6). Among these participants, P5 and P6 were considered professionals due to their formal training in design. None of the participants had prior experience creating found object drawings, but they had seen such works on online platforms such as Instagram and TikTok, and expressed strong interest in this art form.

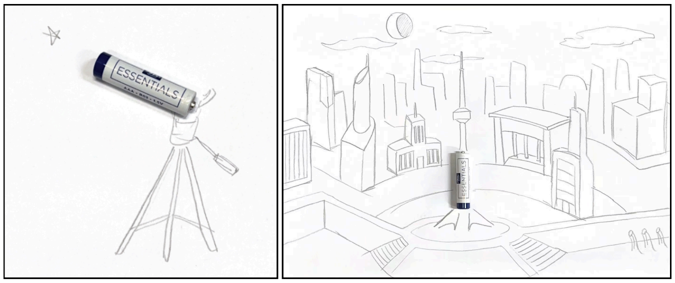


Fig. 2. Two representative drawings in the formative study: the left one (from P1) shows a telescope (object-level), and the right one (from P6) shows a night scene of a modern city (scene-level).

3.2 Procedure

We conducted a 40-minute in-person session with each participant, consisting of a task explanation (5 minutes), a creative task (15 minutes), and a semi-structured interview (20 minutes). The participants were asked to create multiple found object drawings using a battery as the found object. We chose batteries because they are common everyday objects with a simple cylindrical shape, making them accessible for creative reinterpretation. We believed this would encourage active participation, enabling us to observe the participants' creative process and collect valuable insights. During the task, the participants were required to think aloud and explain their motivation for each step. The follow-up interview focused on their creative process and challenges encountered. Each participant received a 7 USD coupon as compensation, and sessions were recorded with consent for later analysis.

3.3 Findings

3.3.1 Creative Process

Our observations revealed a three-step creative process in found object drawing: observing and thinking, searching for reference images, and drawing. The initial observation stage is critical as it requires the participants to reinterpret everyday objects into new visual concepts, though some participants (e.g., P2) found this challenging. We noticed that the participants rarely used reference images during the first stage, possibly due to limitations in existing search tools (e.g. Google Search). Traditional image search methods, whether text-based or image-based, are not well-suited for found object drawing: image-based searches focus on visual similarity rather than creative transformation, while text-based searches require predetermined concepts. As a result, the participants relied more heavily on their imagination. Most participants (P1–5) used reference images primarily for developing ideas or refining drawings in later stages, while P6 relied less on reference images throughout the process, as he had extensive drawing experience.

3.3.2 Challenges and Needs

C1: associative thinking. Most participants were able to come up with associations from the battery (number of results: M = 5.2, SD = 2.39), and those (P5 and P6) having a design background produced the most, with 6 and 9 results, respectively. Other participants (P1, P3, and P4) initially suggested a few simple ideas and found it challenging to continue generating more, while P2 gave up after coming up with a vague idea. The participants with a design background (P5 and P6) were able to generate various ideas efficiently in the beginning, but gradually experienced a decline in their ideation pace. To overcome this, they used auxiliary methods to facilitate further brainstorming, such as observing their surroundings or randomly searching for images for inspiration but with little success. P5 said, *"The battery can be viewed as a cylinder, and any cylindrical object in daily life can be considered."* Similar statements were also made by P1, P3, P4, and P6, and they indicated that geometric features, rather than other attributes such as color or texture, served as their primary source for making associations. However, even for such a simple shape, it remained challenging for them to come up with satisfying ideas. Several participants (P1, P3–5) mentioned that while it was easy to associate various cylindrical objects from everyday life, they preferred ideas that were unique and hard to conceive, which placed high demands on their associative abilities. P2 mentioned that after starting the experiment, he focused on the battery itself, which led him to fall into a fixed mindset, making it difficult to notice other more apparent characteristics. All participants expected a solution to provide them with numerous unique ideas.

C2: visual effect. We categorized the participants' drawing results into two types: object level and scene level, as shown in Figure 2. The categorization was based on the complexity and context of the drawings. At the object level, several participants (P1, P3, and P4) focused on drawing objects with their distinctive features. For example, P1 imagined the battery as the muzzle of a gun and attempted to draw a futuristic-looking gun, while P3 envisioned the battery as a candle and created an elaborate birthday cake. At the scene level, some participants (P2, P5, and P6) incorporated additional elements to depict a broader contextual scene. For instance, P5 envisioned the battery as a locomotive and illustrated a steam train traveling through a forest, while P6 portrayed a skyscraper within a night cityscape (Figure 2 (Right)). The participants who preferred object-level drawings enjoyed portraying the characteristics and details of individual objects. P4 explained, *"I tend to focus on the target object, making it unique and interesting."* Meanwhile, the participants who preferred scene-level drawings explained that this approach allowed them to add more creative content and context, making their visuals feel richer and more complete. P5 said, *"I imagined the battery as a locomotive, and adding the forest and train tracks made the drawing more complete and visually interesting."* Similarly, P6 mentioned, *"The skyscraper alone felt too plain, so I added the cityscape to make it more impressive."* The participants also emphasized the importance of considering both the overall mood and visual elements, with P5 and P6 noting that they would contemplate the desired atmosphere before adding specific details. However, conceptualizing creative visuals was challenging due to imagination constraints, all participants reported a desire to gain inspiration from different perspectives to generate more creative visuals.

C3: reference image. Once they had a rough idea, the participants would use Google Search to find reference images to gain more inspiration or concrete their ideas. However, during this process, they often encountered issues such as mismatched forms, layouts, or styles in the reference images they found. This challenge was particularly notable for the participants with limited drawing skills (P1–4). P1 mentioned, *"I heavily relied on reference images for drawing, but I felt frustrated that they never fully meet my requirements."* They also found it challenging to use reference images for incorporating found objects. P3 said, *"Due to the differences in shape and size between found objects and the elements in reference images, it was hard to imagine how to integrate them into the drawing."* Converting reference images into simple line drawings posed another obstacle, as they struggled to outline and structure the images on paper. The participants with a design background (P5 and P6) still needed the help of suitable reference images. They stated that these references helped them further develop their ideas and depict details, making the final drawings more vivid. All participants agreed that having proper and personalized reference images that better matched their creative needs would significantly improve the creative process.

3.4 Design Goals

Based on the findings above, we distill three design goals for a novel system facilitating found object drawing creation:

DG1: support broad and diverse associative thinking. The system should assist users in generating diverse creative associations based on the found object and allow them to flexibly control the direction of their exploration, enabling them to break away from fixed mindsets (**C1**).

DG2: support multilevel exploration of visual effect. The system should assist users in developing both object-level and scene-level visuals while allowing them to convey creative expression (e.g. moods) in their works (**C2**).

DG3: support personalized reference generation. The system should assist users in generating reference images that align with their creative ideas and provide guidance for them to create drawings using the found object (**C3**).

4 GenFODrawing System

In this section, we begin by presenting the framework of the system in Section 4.1, followed by an overview of its user interface in Section 4.2, and conclude with details about the implementation in Section 4.3.

4.1 Framework

Based on the findings from the formative study, we propose *GenFODrawing*, a novel system that integrates multiple generative models, including VLM, LLM, T2I, and I2I models, to assist users in creating found object drawings. The framework of our system is shown in Figure 3. Given a user-specified photo, users can first select an object of interest (the found object) by clicking on it, which is driven

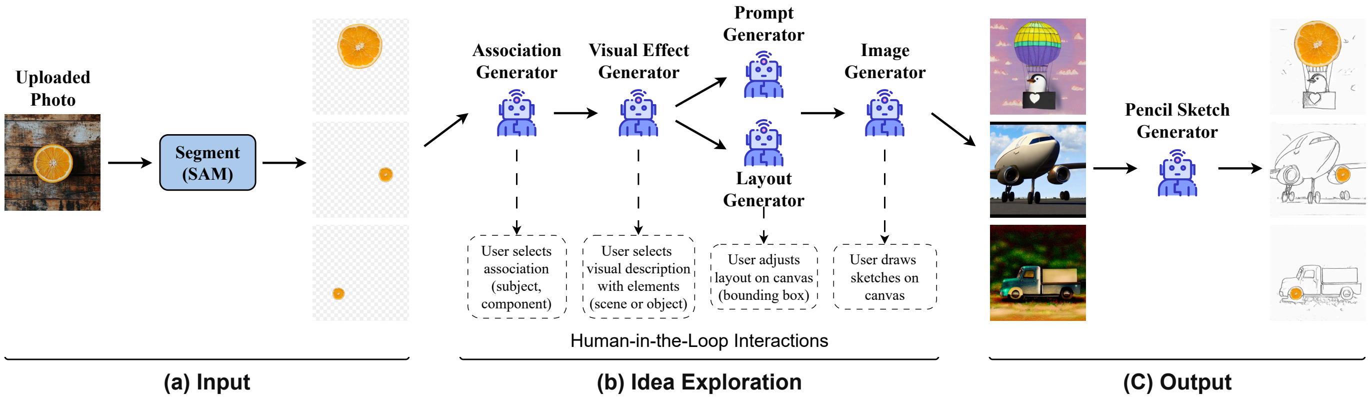

Uploaded Photo
Segment (SAM)
Association Generator
Visual Effect Generator
Prompt Generator
Layout Generator
Image Generator
Pencil Sketch Generator
User selects association (subject, component)
User selects visual description with elements (scene or object)
User adjusts layout on canvas (bounding box)
User draws sketches on canvas
Human-in-the-Loop Interactions
(a) Input
(b) Idea Exploration
(C) Output

Fig. 3. The *GenFODrawing* framework consists of three stages: (a) Input, where users upload a photo and select a found object through segmentation. (b) Idea Exploration, where the system leverages a series of generative modules (association generator, visual effect generator, layout generator, prompt generator, and image generator) to collaboratively generate creative visual references. During this stage, users actively guide the process by selecting associations, refining visual descriptions, adjusting layout compositions, and adding sketches. This human-in-the-loop workflow ensures that the generated references align with user intent. The dashed labels below indicate the user-guided decision points. (c) Output, where the final reference image is converted into a pencil sketch with the overlaid found object, supporting users in manual drawing.

by a segmentation model. Users can freely translate, rotate, and scale the selected found object on the canvas as the beginning of idea exploration. During idea exploration, we adopt a human-in-the-loop interaction design, where five generators collaborate to help users transform found objects into personalized reference images. The association generator provides creative, shape-based suggestions for the found object while considering context such as drawable regions and object placement. Users can guide association generation by setting themes or selecting example associations, enabling diverse and controllable idea exploration (**DG1**). Once an association is chosen, the visual effect generator recommends expressive scene- or object-level descriptions. Users can further adjust the mood or modify elements and visual descriptions to better align the results with their intended outcomes (**DG2**). The layout generator then suggests layouts for the elements in the selected visual description. Users can freely adjust the bounding boxes of elements and provide sketches on the canvas to guide image synthesis. Prompts for image generation are automatically constructed by the prompt generator, which incorporates both the selected visual description and user-specified style (see Section 4.3 for details). Finally, users can convert the generated images into pencil sketches and overlay the found object onto them, providing clear visual guidance for drawing on paper (**DG3**).

4.2 User Interface

As illustrated in Figure 4, *GenFODrawing* consists of three panels to facilitate the creation of found object drawings for users. Below, we will introduce each panel in detail.

4.2.1 Drawing Panel

Before idea generation, the user needs to upload a photo containing a found object of interest to our system. The drawing panel (Figure 4 (Middle)) displays the uploaded image and allows the user to directly select the found object by segmentation. The user can freely transform the segmented found object on the drawing canvas to control its position, rotation, and scale. This allows the user not only to guide the generation of creative associations, but also to specify the desired placement of the found object in the final composition. It also provides auxiliary tools like a brush and eraser, enabling the user to sketch his/her envisioned images on the drawing canvas. Additionally, the drawing panel can automatically capture the outer contour of the found object, allowing the user to quickly use the partial or entire shape of the object. The user can also click the "Generate Layout" button to automatically generate a layout based on the visual description and elements obtained from the idea generation panel. Each element's bounding box can then be freely adjusted in terms of position and scale.

4.2.2 Idea Generation Panel

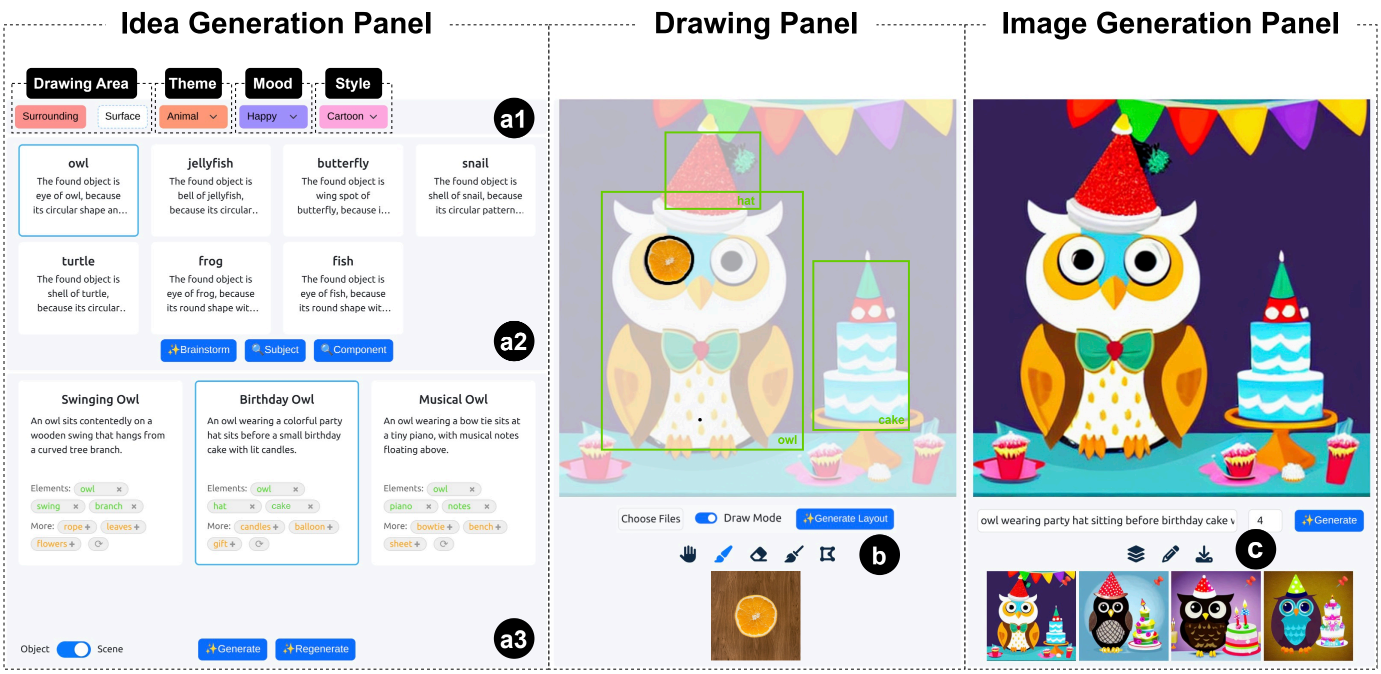
The idea generation panel (Figure 4 (Left)) consists of three parts: the user-specified options, the association recommendations, and the visual description recommendations. To better support the user in idea generation, we designed several features (Figure 4a1) with the following rationales:

Drawing area. Found object drawings can be created by adding elements around or on the found object itself. To support these two fundamental approaches, we designed a drawing area configuration feature that allows the user to specify his/her intended drawing space explicitly. The user can choose between two options: the "Surrounding" button to draw around the found object, or the "Surface" button to draw on its surface.

Theme. When creating found object drawings, people need to make meaningful associations within specific contexts. To assist with this process, we introduced a theme configuration feature that narrows down possible associations to specific categories, such as animal, plant, vehicle, etc.

Mood. Found object drawings can be more expressive when emotional aspects are incorporated. To help the user achieve his/her desired emotional effect, we introduced a mood configuration feature. The user can specify desired moods, such as happy, surprised, frustrated, etc.

Style. Different artistic styles can significantly influence the visual impact of the final drawing. To give the user control over the visual appearance of the work, we introduced a



Surrounding | Surface | Animal | Happy | Cartoon

owl
The found object is eye of owl, because its circular shape an…

jellyfish
The found object is bell of jellyfish, because its circular..

butterfly
The found object is wing spot of butterfly, because i…

snail
The found object is shell of snail, because its circular pattern…

turtle
The found object is shell of turtle, because its circular..

frog
The found object is eye of frog, because its round shape wit…

fish
The found object is eye of fish, because its round shape wit…

Swinging Owl
An owl sits contentedly on a wooden swing that hangs from a curved tree branch.
Elements: owl × swing × branch ×
More: rope + leaves + flowers + ⟳

Birthday Owl
An owl wearing a colorful party hat sits before a small birthday cake with lit candles.
Elements: owl × hat × cake ×
More: candles + balloon + gift + ⟳

Musical Owl
An owl wearing a bow tie sits at a tiny piano, with musical notes floating above.
Elements: owl × piano × notes ×
More: bowtie + bench + sheet + ⟳

Object ● Scene

owl wearing party hat sitting before birthday cake v | 4

Fig. 4. The user interface of *GenFODrawing* consists of three panels: the idea generation panel providing (a1) user-specified options, (a2) association recommendations, and (a3) visual description recommendations; the drawing panel offering a canvas and (b) several tools for found object selection, manipulation, sketch creation, and contour extraction; and the image generation panel featuring a main canvas displaying generated reference images, a gallery of generation results, and (c) functional buttons to overlay found objects onto reference images and convert results to pencil sketches for previewing final outcomes.

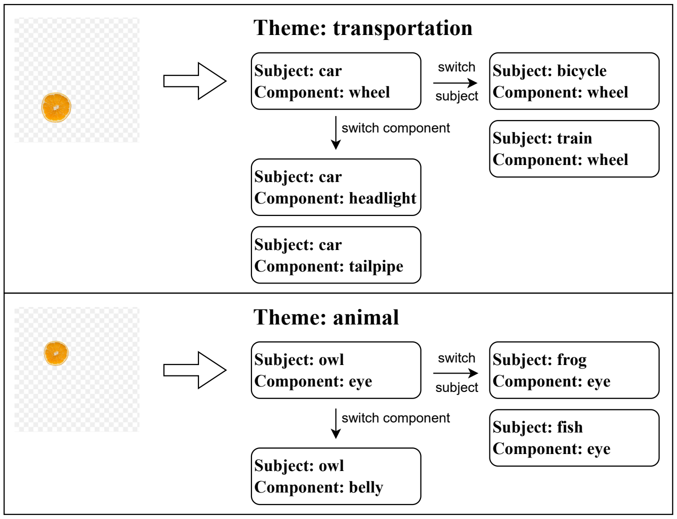


Fig. 5. Two examples of interactively exploring diverse associations by specifying theme constraints and switching subjects or components.

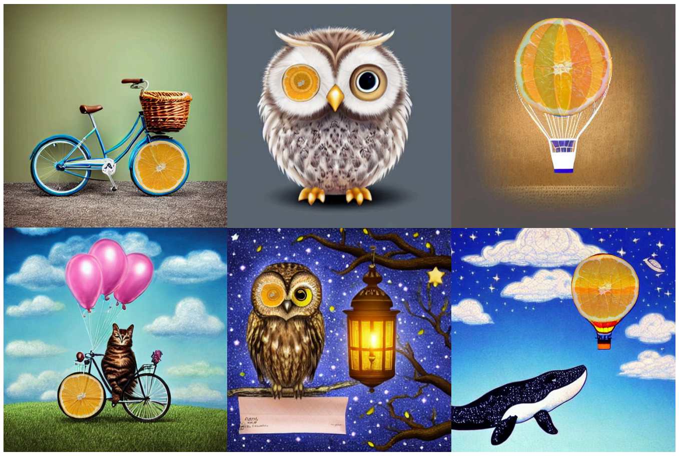


Fig. 6. Different object- (the up row) and scene- (the bottom row) level reference images recommended by *GenFODrawing*, with the found object (orange slice) overlaid on them.

style configuration feature. The user can specify the artistic style, such as realistic, illustration, cartoon, etc.

For theme, mood, and style configurations, the user can choose from the preset options or input his/her custom values. Once the user has specified the theme, he/her can click the "Brainstorm" button to ask the system to recommend associations with the found object (Figure 4a2). Recommendations are presented as cards, each corresponding to an implicit (subject, component) pair, which indicates the functional component the found object can serve in a particular subject. The card's title displays the subject, while the explanation is shown in the format: "The found object is <component> of <subject>, because...". The user can hover over cards to view the complete explanation. When selecting an association of interest, the user can further explore similar associations using the "Subject" and "Component" buttons, which generate new recommendations by switching either the subject or the component, as shown in Figure 5.

After selecting an association of interest, the user can explore visual descriptions that incorporate the subject, which are displayed in the lower part of the panel (Figure 4a3). Each recommendation is presented as a card containing a title and a detailed description. The system automatically identifies the main elements present in the current description (green tags) and also suggests additional elements (orange tags). The user can freely add, delete, or edit elements in the list to better fit his/her creative ideas. By clicking the "Regenerate" button, the user can update the visual description according to his/her customized element selection, allowing for broader exploration of different image possibilities. The mood specified by the user will also be considered when the system generates visual descriptions. At the bottom left, a toggle button enables the user to switch between scene-level and object-level recommendations. As shown in Figure 6, the "Scene" mode provides suggestions for overall scene descriptions, while the "Object" mode focuses on the appearance of individual objects. At the bottom center, the "Generate" and "Regenerate" buttons allow the user to create or update visual descriptions based on his/her current selections. This design encourages the user to flexibly explore a wide range of visual effects and quickly iterate on his/her creative concepts.

4.2.3 Image Generation Panel

The image generation panel (Figure 4 (Right)) can generate and display reference images based on the sketch drawn on the drawing panel and the provided text prompts. The text prompt can either be entered by the user or, if not specified, is automatically recommended by the system, taking into account the user's style preferences. This panel also offers a feature that allows the found object to be overlaid on the current reference image, making it easier for the user to assess whether it matches the found object (Figure 6). Additionally, it provides a function to convert colored images into pencil sketches, helping the user better perceive the final visual outputs. The user can freely switch between different generated results listed below to examine these features on each generated image.

4.3 Implementation

GenFODrawing is a web-based system composed of a ReactJS front-end and a Python Flask server as its back-end, deployed on a machine with an NVIDIA RTX 4090. To select a found object of interest from the uploaded photo, we use the *Segment Anything Model* [47] to allow users to click on object areas of interest for segmentation. Given GPT-4o's [48] powerful multimodal capabilities, we employ it as both our VLM and LLM underlying models in the system. All prompts used in our system are provided in the supplementary materials.

For association recommendation, we design a multi-step prompting strategy that combines in-context learning and chain-of-thought reasoning. The prompt provides example associations and explicit instructions to guide the model in generating (subject, component) pairs that are creative yet recognizable, and non-duplicative (using a historical record). User-specified examples or target themes are embedded in the prompt to support relevance and user control. Each output is returned in a standardized JSON format with concise explanations to enhance transparency and usability. After generating candidate associations, we apply a spatial reasoning filter via an additional prompt step to ensure contextual appropriateness. Specifically, the model is prompted to examine each association and reason whether the current orientation and size of the found object realistically fit the expected component within the subject, and whether its placement would allow for a natural and recognizable composition. Associations that may lead to awkward, unnatural, or unrecognizable results (e.g., mismatched rotation, improper scale, or interference from non-drawable regions) are filtered out. This spatial reasoning step helps ensure that the recommended associations are not only semantically plausible but also visually feasible for subsequent drawing tasks.

For visual description recommendation, we first prompt the generator to produce different scenes or object variations that reflect the specified mood. Then, we use the text encoder of OpenCLIP [49] ('ViT-G-14') to encode each candidate description d_i into a text embedding e_i. These embeddings capture the semantic meaning of the input text, enabling us to measure semantic similarity between descriptions via cosine similarity. We compute the pairwise cosine similarity between all embeddings $E = \{e_1, e_2, \ldots, e_n\}$ and seek a maximal subset $D' \subset D$ such that:

$$\cos(e_i, e_j) < \delta, \quad \forall\, d_i, d_j \in D',\, i \neq j \tag{1}$$

where $\cos(e_i, e_j)$ denotes the cosine similarity between the embeddings of descriptions d_i and d_j, and $\delta = 0.6$ is the similarity threshold used in our implementation. This constraint ensures that all descriptions retained in D' are sufficiently diverse in semantics, avoiding similar suggestions.

For T2I prompt recommendation, we use a template of `[image info, style tags, quality tags]` to guide the generator in producing high-quality prompts that align with the selected visual description and desired style. The prompt explicitly includes all elements mentioned in the selected visual description. For layout recommendation, we further instruct the generator to provide bounding boxes for all elements included in the selected visual description, using the format `[x_min, y_min, x_max, y_max]`.

For image generation, we employ a combination of Stable Diffusion v1.5 [50] and a ControlNet model [51] fine-tuned on a partial sketch dataset, since during the early stages of user exploration, users often provide only minimal sketch input (e.g., the contour of the found object) as guidance. To enable users to freely explore different placements of the found object, we adopt a box-conditioned generation approach similar to [52]. Specifically, at each denoising step t, given the noised sample $\mathbf{x}_t$, we update it via gradient descent to minimize both cross-attention and self-attention losses:

$$\hat{\mathbf{x}}_t \leftarrow \mathbf{x}_t - \alpha \nabla_{\mathbf{x}_t} \left(\mathcal{L}_{CA} + \mathcal{L}_{SA} \right), \tag{2}$$

where α is the step size, $\mathcal{L}_{CA}$ is the cross-attention loss, and $\mathcal{L}_{SA}$ is the self-attention loss.

To obtain the cross-attention map, we first perform a forward pass of the diffusion model and ControlNet with the current noised latent x_t, the text prompt, and the sketch image as inputs. We extract the cross-attention maps from all layers and heads at the 16×16 spatial resolution and average them to produce a single aggregated attention tensor of shape $16 \times 16 \times n$, where n is the number of text tokens. We then discard the attention map of the special start-of-text token $\langle \mathrm{sot} \rangle$ and re-normalize the remaining attention values along the token dimension using $\mathrm{softmax}$. As a result, each

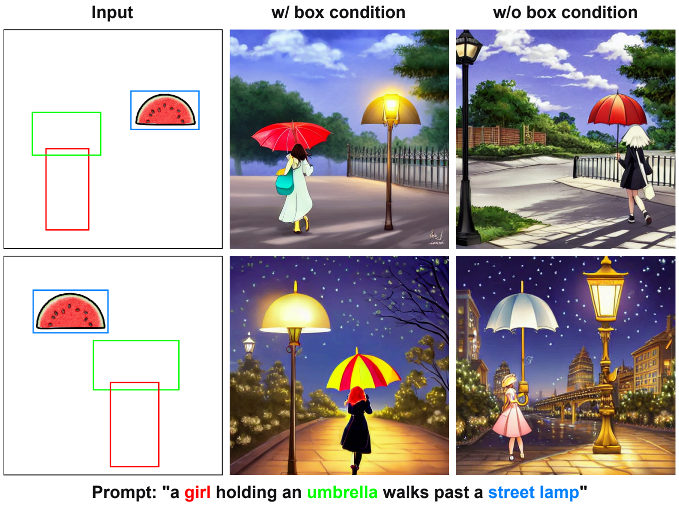

Input
w/ box condition
w/o box condition
Prompt: "a girl holding an umbrella walks past a street lamp"

Fig. 7. The figure shows two examples of using the object contour as a partial sketch input under two conditions. Without box conditioning, users can only express their intent through the sketch, and the generated content may not match their intent (e.g., in the third column, the found object is generated as an umbrella instead of a street lamp). With box conditioning, users can further control the generation with bounding boxes, resulting in outputs that better align with their creative intent (the second column). For each row, we use the same seed.

spatial location (i, j) in $\mathcal{A}^t$ indicates the probability of each token being present in the corresponding image patch. The cross-attention loss is computed separately for the foreground (inside the bounding box) and background (outside the bounding box) regions. The overall cross-attention loss is defined as:

$$\mathcal{L}_{CA} = \mathcal{L}_{fg} + \mathcal{L}_{bg}. \tag{3}$$

The foreground term encourages high attention inside the corresponding bounding box and is defined as:

$$\mathcal{L}_{fg} = \frac{1}{N}\sum_{i=1}^{N}\left[1 - \text{mean}\left(\text{Top}_p\left(\mathcal{A}_i^t \odot M_i\right)\right)\right], \tag{4}$$

where $\text{Top}_p(x)$ denotes the set of the top $p\%$ largest values of x (we set $p = 80$ in our implementation). Focusing only on the top $p\%$ activations avoids imposing constraints on all attention values, which could harm image fidelity. $\mathcal{A}_i^t$ is the cross-attention maps for the i-th element at denoising step t, M_i is the binary mask for the bounding box of the i-th element, and N is the number of elements. The background term is computed similarly, penalizing high attention values outside the corresponding bounding box:

$$\mathcal{L}_{bg} = \frac{1}{N}\sum_{i=1}^{N}\text{mean}\left(\text{Top}_p\left(\mathcal{A}_i^t \odot (1 - M_i)\right)\right). \tag{5}$$

Overall, the foreground and background terms concentrate token-to-patch attention inside the corresponding bounding boxes and suppress it elsewhere, thus improving the consistency between the generated objects and the given spatial layout.

The self-attention maps are obtained in a similar way to the cross-attention maps, yielding tensors of shape $16 \times 16 \times 256$. Each spatial location (i, j) in $\mathcal{S}^t$ indicates how much information the current patch draws from every other patch in the image. The self-attention loss $\mathcal{L}_{SA}$ encourages the generated content to remain faithful to the region defined by the corresponding bounding box:

$$\mathcal{L}_{SA} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|M_i|}\sum_{p\in M_i}\text{mean}\left(\mathcal{S}_p^t \odot (1 - M_i)\right), \tag{6}$$

where $|M_i|$ is the number of patches inside the mask, and $\mathcal{S}_p^t$ is the self-attention maps at step t for patch p. By penalizing strong self-attention connections between masked and unmasked patches, this loss limits information leakage from the object region to patches outside the bounding box, reducing the risk of propagating object-specific features to unrelated areas. As shown in Figure 7, directly generating from limited sketch input can result in inconsistency with user intent. By conditioning on both sketches and bounding boxes, we achieve more controllable and precise image synthesis. This approach allows the found object to be flexibly positioned and integrated within the generated image. We also filter out images with low text-image CLIP scores (below 0.4 in our implementation) to ensure the quality and relevance of the results. The reference images can then be transferred to pencil sketches using a style transfer model [53] with a U-Net structure.

5 Evaluation

We first conducted a within-subjects comparative study to investigate the effectiveness of *GenFODrawing*. Since no existing system supports the creation of found object drawings, we compared *GenFODrawing* to a baseline condition in which participants completed the creative tasks without access to our system. In the baseline condition, participants were free to use any tools or methods they normally rely on, such as ChatGPT, Google Search, or their own imagination, allowing us to establish a realistic "without system support" scenario. This design enables a direct comparison between the experience with our structured, controllable creative support system and the experience without such targeted assistance. After the comparative study, we conducted an open-ended study to further evaluate the usability and expressiveness of *GenFODrawing*.

5.1 Participants

We recruited twelve participants (9 males and 3 females, P1 to P12), with ages ranging from 25 to 31 (M = 27.75, SD = 1.66). They are postgraduate students with diverse backgrounds, from engineering to art. The demographic information is provided in the supplementary materials. All participants had seen found object drawings before, but none had prior experience in creating such works.

5.2 Tasks, Procedure, and Measures

The user study consisted of two phases. In the first phase, a comparative experiment was conducted where the participants created found object drawings based on four given objects: two different leaves and two different clips, as shown in Figure 8. These objects were divided into two groups, with the objects in Figure 8(a) and (b) forming one group, and those in Figure 8(c) and (d) forming the other group. For each participant, one group was assigned to the baseline

TABLE 1
The CSI and controllability questionnaires used in the comparative study.

Q1	I would be happy to use this tool on a regular basis.
Q2	It was easy for me to explore many different ideas, options, designs, or outcomes, using this tool.
Q3	I was able to be very creative while doing the activity inside this tool.
Q4	I became so absorbed in the activity that I forgot about the tool that I was using.
Q5	What I was able to produce was worth the effort I had to exert to produce it.
Q6	I am able to adjust the direction of idea exploration quickly and efficiently according to my creative intention.
Q7	I am able to precisely control and iteratively refine the generated results to better fit my creative intention.
Q8	The system provides a clear and structured process for me to explore a variety of creative ideas.

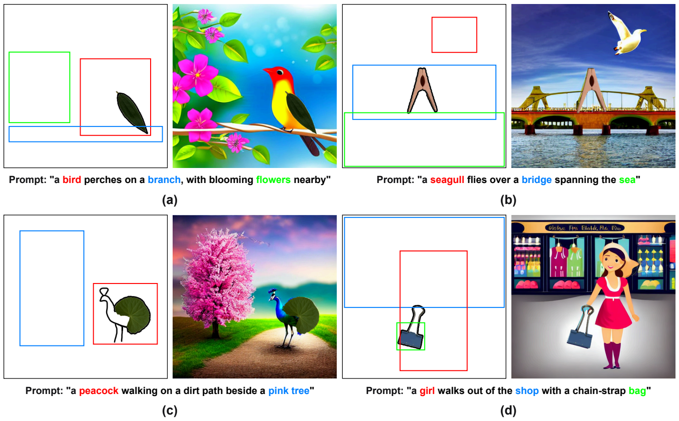

Prompt: "a bird perches on a branch, with blooming flowers nearby"
(a)
Prompt: "a seagull flies over a bridge spanning the sea"
(b)
Prompt: "a peacock walking on a dirt path beside a pink tree"
(c)
Prompt: "a girl walks out of the shop with a chain-strap bag"
(d)

Fig. 8. The figure shows some results created by P12 (a), P3 (b), P4 (c), P6 (d) using *GenFODrawing* in the comparative study. The first and third columns show the input sketches along with the specified bounding boxes, while the second and fourth columns show the system's outputs, with the found object overlaid on the images.

condition and the other to the *GenFODrawing* condition, and they were required to create one drawing for each object in both conditions, for a total of four drawings per participant. The assignment of object groups to conditions was counterbalanced across participants to control for order effects, and the sequence of drawing tasks was randomized within each group. After each condition, the participants filled out a questionnaire. After completing both conditions, they engaged in a 15-minute interview. In the second phase, we conducted an open-ended study that allowed the participants to freely choose any nearby object or its shadow to create a found object drawing. They also filled out a questionnaire after completing the open-ended task. The entire study took an average of approximately two hours per participant, and the participants were compensated with a coupon of around 15 USD upon completion.

For measurement purposes, we employed a Creativity Support Index (CSI) [54] questionnaire and a 7-point Likert scale questionnaire (see Table 1) to respectively assess creativity support and controllability in the comparative study. Additionally, we used NASA-TLX [55] and System Usability Scale (SUS) [56] questionnaires to evaluate workload and usability of our system in the open-ended study, respectively. The collected data were analyzed using the Wilcoxon signed-rank test.

5.3 Results

Throughout the experiment under the baseline condition, all participants chose to primarily utilize ChatGPT (powered by GPT-4o), and most of them (except P2) also used its integrated image generation capabilities. Additionally, a few participants used Google Search as a supplementary source. We analyzed the collected data and interview records. In this section, we will report the findings.

5.3.1 Quantitative Survey Results

The statistical results of CSI and controllability questionnaires are shown in Table 2. Regarding the Creativity Support Index (CSI) scores, our system received higher scores than the baseline across all dimensions, including Enjoyment (M = 5.33 vs. 4.08), Exploration (M = 6.00 vs. 3.75), Expressiveness (M = 5.75 vs. 3.50), Immersion (M = 5.33 vs. 3.17), and Results Worth Effort (M = 5.58 vs. 4.58). All differences were statistically significant ($p < 0.05$). These findings indicate that our system provides a significantly enhanced creative experience, particularly in terms of supporting user enjoyment, exploration, expressiveness, and immersion. For controllability, users felt significantly better able to adjust the direction of idea exploration (M = 5.83 vs. 3.42, $p = 0.005$), to precisely control and iteratively refine the generated results (M = 5.83 vs. 3.17, $p = 0.004$), and to benefit from a clear and structured creative process (M = 5.75 vs. 3.08, $p = 0.003$). These results highlight the advantages of our system in supporting user creativity and providing enhanced control over the creative process, with particularly strong improvements in user exploration, expressiveness, and controllability.

5.3.2 Qualitative Feedback

Creative Inspiration, Exploration, and Accessibility. The participants consistently reported that our system provided more effective creative inspiration compared to the baseline. The structured, stepwise guidance and multimodal recommendations (both textual and visual) helped users overcome creative blocks and discover new ideas, especially when they *"had no clear direction at the start"* (P4). Some participants also noted that the system's ability to specify moods and adjust elements enabled them to generate a broader range of ideas and visual possibilities. As P1 described, *"The system's suggestions sometimes led me to ideas I wouldn't have considered on my own."* And P2 noted, *"The system's structured design and visual feedback are its strengths. It guides me step by step, which lowers the mental demand and makes it easy to focus on each small part of the process."* By contrast, in the baseline condition (ChatGPT), most participants (P1-3, P5-12) found it necessary to prompt the model multiple times, either to request ideas or to generate images based on the object. This process often resulted in increased cognitive load, as users had to repeatedly construct, edit, and rephrase prompts to steer the output toward their intended outcomes, as shown in Figure 9 (Baseline). As P10 commented, *"I am not good at prompt engineering, and sometimes I don't know how to*

TABLE 2
The statistical results of CSI and controllability questionnaires. ($*$: $p < 0.05$ and $**$: $p < 0.01$).

		GenFODrawing		**Baseline**		Statistics	
		mean	std	mean	std	p-value	Sig.
Creativity Support	Enjoyment	5.33	1.37	4.08	1.38	.013	$*$
	Exploration	6.00	0.95	3.75	1.36	.003	$**$
	Expressiveness	5.75	1.06	3.50	1.57	.003	$**$
	Immersion	5.33	1.07	3.17	0.94	.001	$**$
	Results Worth Effort	5.58	1.00	4.58	1.00	.028	$*$
Controllability	Direction Adjustment	5.83	0.83	3.42	1.16	.005	$**$
	Result Refinement	5.83	0.94	3.17	1.34	.004	$**$
	Structured Process	5.75	0.62	3.08	1.38	.003	$**$

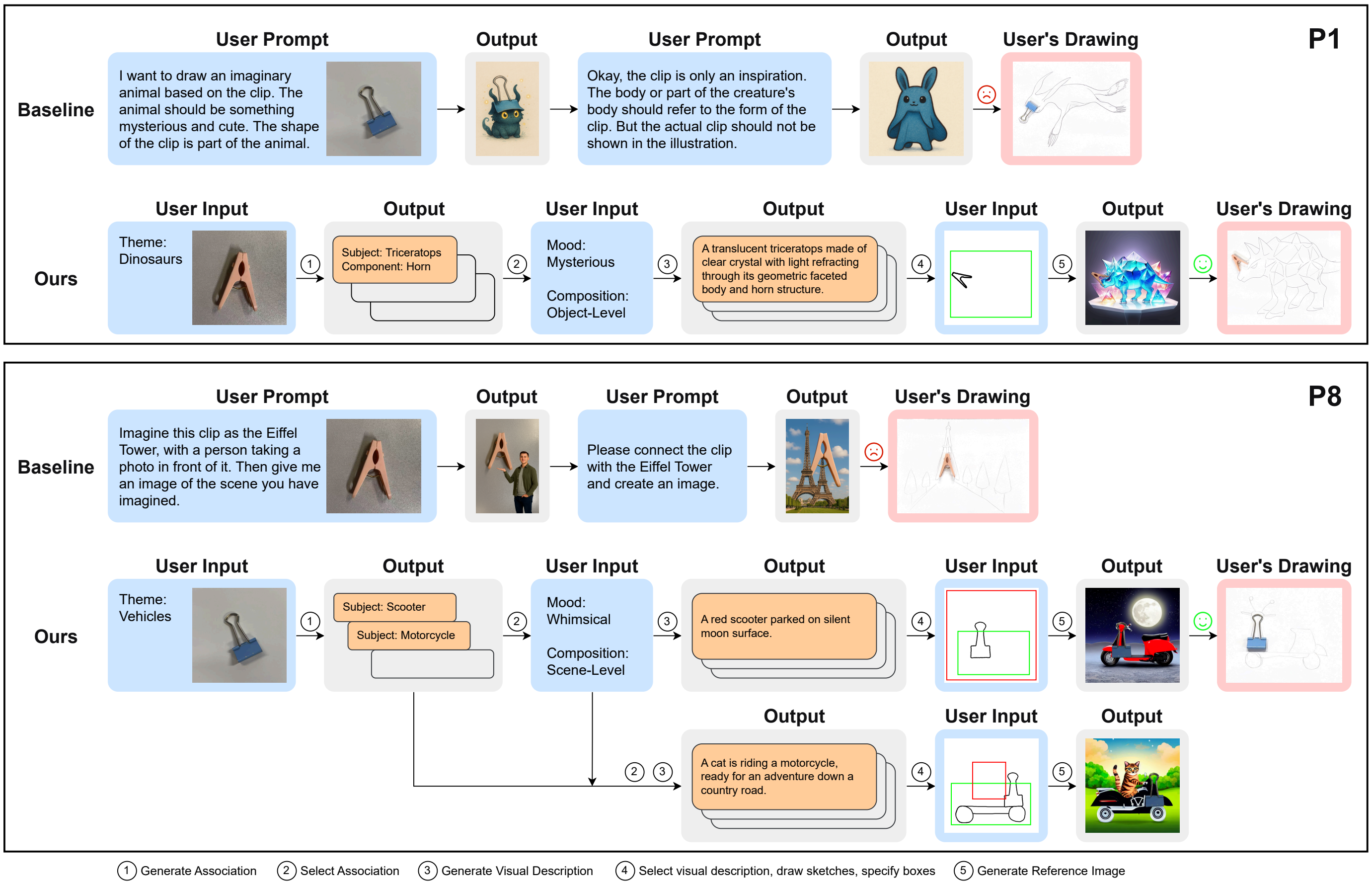

P1
Baseline
User Prompt
I want to draw an imaginary animal based on the clip. The animal should be something mysterious and cute. The shape of the clip is part of the animal.
Output
User Prompt
Okay, the clip is only an inspiration. The body or part of the creature's body should refer to the form of the clip. But the actual clip should not be shown in the illustration.
Output
User's Drawing
Ours
User Input
Theme: Dinosaurs
Output
Subject: Triceratops
Component: Horn
User Input
Mood: Mysterious
Composition: Object-Level
Output
A translucent triceratops made of clear crystal with light refracting through its geometric faceted body and horn structure.
User Input
Output
User's Drawing
P8
Baseline
User Prompt
Imagine this clip as the Eiffel Tower, with a person taking a photo in front of it. Then give me an image of the scene you have imagined.
Output
User Prompt
Please connect the clip with the Eiffel Tower and create an image.
Output
User's Drawing
Ours
User Input
Theme: Vehicles
Output
Subject: Scooter
Subject: Motorcycle
User Input
Mood: Whimsical
Composition: Scene-Level
Output
A red scooter parked on silent moon surface.
User Input
Output
User's Drawing
Output
A cat is riding a motorcycle, ready for an adventure down a country road.
User Input
Output
1 2 3 4 5
1 Generate Association
2 Select Association
3 Generate Visual Description
4 Select visual description, draw sketches, specify boxes
5 Generate Reference Image

Fig. 9. The figure shows part of the creative process for P1 and P8 under both the baseline (ChatGPT) and our system conditions in the comparative study. Due to space limitations, only excerpts of the ChatGPT conversation are shown. When using ChatGPT, participants engaged in a linear, prompt-by-prompt exploration, repeatedly revising prompts to approach their desired outcomes. The outputs were unpredictable and often misaligned with users' creative intent, requiring extra effort and imagination to adapt the results. In contrast, our system guides users through a structured exploration and enables more controllable image generation, making it easier to explore different ideas and achieve their creative goals.

prompt ChatGPT to get the result I want." Moreover, the linear conversation workflow of ChatGPT made it cumbersome to revisit and compare alternative ideas, or to flexibly switch between different creative directions. As a result, the participants reported it was easy to get "stuck" on a single idea and lose track of alternative directions, making it harder to systematically try out different creative possibilities. As P6 remarked, *"With the first system (GenFODrawing), I can always see all the ideas and go back to try another. With the second one (ChatGPT), the conversation keeps moving forward, so I usually just go with the first idea that works."* Overall, the participants found that our system's structured, visual workflow and step-by-step guidance not only supported broader creative exploration and helped them avoid idea fixation but also significantly reduced cognitive demands and made the process more accessible, especially for those with less experience in found object drawing and prompt engineering.

Controllability and User Agency. The majority of the participants highlighted the significantly enhanced sense of control and authorship provided by our system compared to the baseline. Features such as changing visual descriptions, sketching, manipulating bounding boxes, and iterative refinement enabled users to directly guide the creative pro-

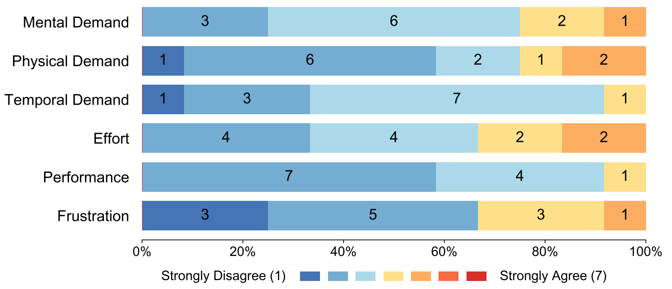


Fig. 10. Distribution of NASA-TLX scores across all six items. Lower scores indicate better perceived user experiences. The complete questions are provided in the supplementary materials.

cess. As P3 noted, *"It's satisfying to adjust the elements visually and see immediate changes, rather than having to constantly figure out how to phrase prompts to steer the model toward what I want."* In contrast, the participants described the baseline ChatGPT workflow as more passive and less controllable: *"I can only use text prompts and hope it understands me, but often the results are not what I imagined. Since it's a black box, I'm not sure whether the problem is with my prompt or if it just doesn't understand me, which can leave me feeling lost"* (P11). Several participants (P1, P6-8, P10-12) mentioned that, in many cases, they simply accepted whatever result ChatGPT produced if it seemed "looked good", even if it did not fully match their intention, because making further adjustments was difficult. As P9 put it, *"I just go along with whatever ChatGPT gives me, instead of really shaping the result myself."* This passive interaction reduced their sense of agency over the final creation. Overall, our system enabled users to actively direct both the creative process and the output, leading to a stronger feeling of authorship and creative ownership. The participants consistently valued the increased transparency, control, and iterative flexibility our interface provided.

User Engagement and Reference Value. The participants held different views on the value of generated images as creative references. Several participants (P1, P6-8, P10-12) noted that ChatGPT's image outputs were often unpredictable. Some results could be used, but others either misunderstood users' prompts or failed to meet users' intentions, making them difficult to use directly as drawing references (see Figure 9 Baseline). As a result, they often needed to experiment with multiple prompts or rely heavily on their own imagination to adapt the generated references and incorporate the found object into their drawing, which was especially challenging for users with less drawing experience. By contrast, our system provided more consistent and controllable outputs. Users could actively influence the generated reference images by customizing the sketch and layout, resulting in richer and more personally relevant visual references (see Figure 9 Ours). This flexibility allowed users to spend more time exploring and iterating on their creative ideas, leading to higher engagement. As P12 said, *"The generated results closely match the sketch and bounding boxes I provided. It's very interesting and makes me want to try more ideas."* P2 and P3 further noted that, since the references generated by our system follow the sketch and layout inputs, they could also use them as guidance for composition when drawing on paper. Notably, our system also provided a function to convert images into pencil sketch outputs, which several participants found helpful for lowering the barrier for beginners. As P8 commented, *"This function makes it much easier for me to start, I can focus on the main structure instead of all the details."* Overall, compared to the baseline, our system provided greater stability and richer reference values, fostering more active and sustained creative engagement.

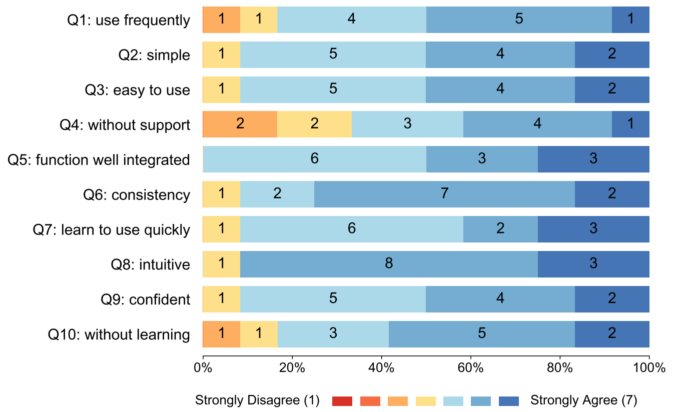


Fig. 11. Distribution of SUS scores across all ten items. Each label represents a key summary of the corresponding SUS question. Higher scores indicate better perceived user experiences. The complete questions are provided in the supplementary materials.

5.3.3 Open-ended Study Results

In the open-ended study phase, the participants were invited to freely select objects and use our system to create found object drawings, aiming to evaluate the system's usability across a wide range of user-chosen inputs. All participants successfully completed the task. They selected a diverse array of found objects, including everyday items (e.g., earpods, hair clips), natural materials (e.g., leaves, petals), and shadows, which demonstrates the system's broad applicability. Users explored different creative directions by adjusting associations, visual descriptions, and sketch overlays, with several participants iteratively refining their work through multiple system features. The distribution of NASA-TLX scores is shown in Figure 10. Overall, users reported relatively low levels of perceived workload when using our system with satisfactory performance. The average scores for Mental Demand (M = 3.08, SD = 0.90), Physical Demand (M = 2.75, SD = 1.29), Temporal Demand (M = 2.67, SD = 0.78), Effort (M = 3.17, SD = 1.11), and Frustration (M = 2.50, SD = 1.38) were all on the lower end of the 7-point scale, indicating that the participants generally found the system easy and comfortable to use. The average Performance score was high (M = 5.50, SD = 0.67), suggesting that users felt satisfied with their outcomes. The distribution of SUS scores is shown in Figure 11. The results indicate generally positive user perceptions. The participants expressed a strong willingness to use the tool frequently (M = 5.33, SD = 1.07) and found it easy to use (M = 5.58, SD = 0.90), with most agreeing that its functions were well integrated (M = 5.75, SD = 0.87) and that they could learn to use it quickly (M = 5.58, SD = 1.00). Confidence in using the tool was also high (M = 5.58, SD = 0.90). In contrast, the participants largely disagreed with statements

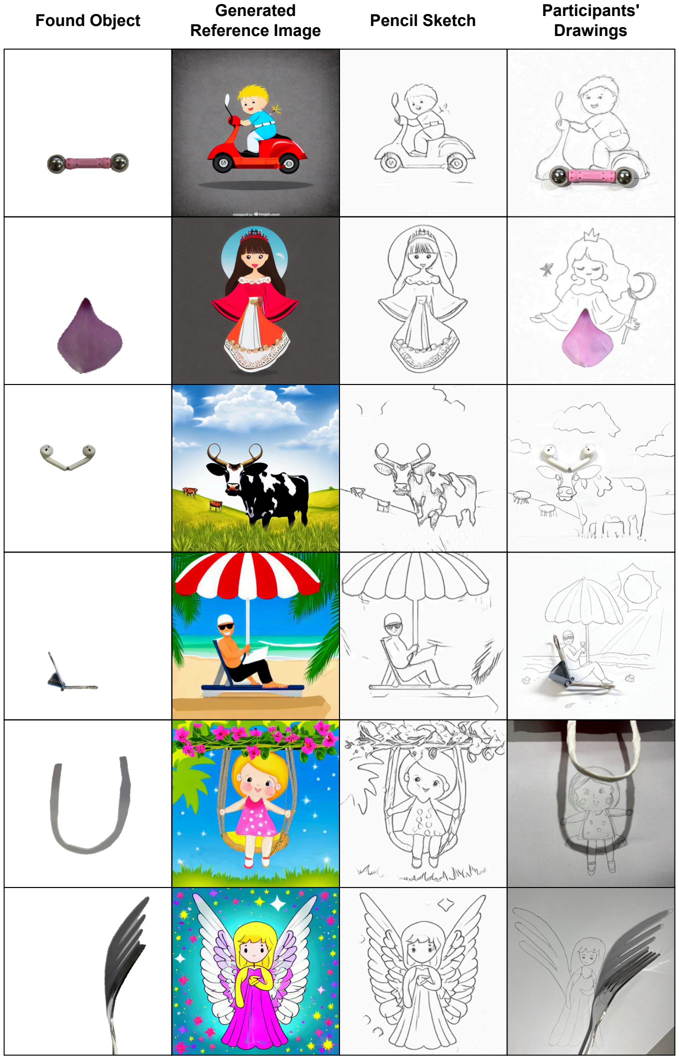


Fig. 12. Selected results created by the participants in the open-ended study. More results can be found in the supplementary materials.

suggesting the tool was unnecessarily complex (M = 2.42, SD = 0.90), inconsistent (M = 2.17, SD = 0.83), or awkward to use (M = 1.92, SD = 0.79). Scores related to the need for technical support (M = 3.00, 1.28) and the amount of learning required (M = 2.50, 1.17) were moderate, indicating a manageable learning curve. These results demonstrate that the system is accessible, well-integrated, and supports a positive overall user experience.

Several selected results are shown in Figure 12. For example, P3 used magnetic sticks and balls as the wheels and pedals of a scooter (the first row); P4 transformed a flower petal into a skirt and added her own creative modifications to the girl's figure instead of directly following the reference image (the second row); P5 used the earpods as the cow's horns (the third row); P6 interpreted an open clip as a beach chair, drawing a man sitting on it (the forth row); P9 used the shadow of a bag handle as a swing and drew a girl swinging on it (the fifth row); and P12 used the shadow of a fork as the angel's wings (the sixth row). Such cases illustrate how the system enabled the participants to reinterpret ordinary objects and their shadows as creative components within their work. Overall, the participants found the system effective for stimulating new ideas and supporting creative expression with a wide range of found objects. These findings demonstrate the system's strong usability while maintaining a low workload.

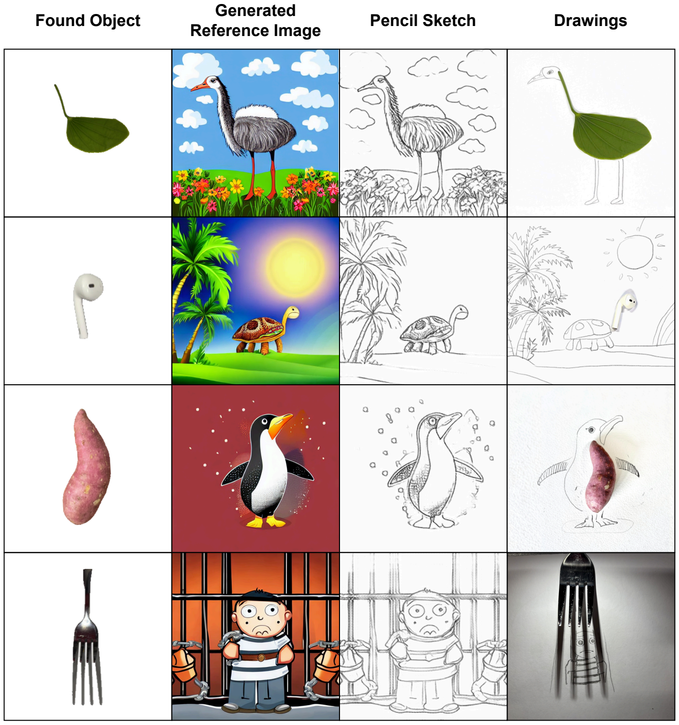


Fig. 13. Some results created by the authors. More results can be found in the supplementary materials.

6 Discussion

6.1 Bridge Physical and Digital Creation through AI

GenFODrawing presents broader implications for integrating AI support into physical-digital creative practices. While generative AI has shown remarkable capabilities in purely digital creation, its application in physical-world creative tasks remains challenging. Our findings reveal several key considerations for bridging this physical-digital divide in creative work. Physical objects and spaces provide natural constraints and affordances that shape creative thinking differently from purely digital environments. While this might seem limiting, our study shows that these physical constraints can actually serve as valuable creative anchors. The challenge lies in helping AI systems understand and leverage these physical contexts, not just as visual inputs, but as meaningful, creative starting points. This suggests a broader need for AI systems that can better interpret and respond to physical creative contexts. The transition between physical observation and digital augmentation also requires careful consideration of interaction design. In *GenFODrawing*, users can transform a physical object into a digital sketch input and preview the final results by overlaying the object onto the generated images, maintaining continuity between the two domains. This suggests that future AI-supported creative tools should offer flexible ways to bridge physical and digital representations. Moreover, these physical constraints should serve not only as creative starting points but also as guiding parameters for AI-generated outputs. This is particularly crucial for novice users who may lack professional training: when AI-generated content respects physical constraints, it becomes more feasible and practical for real-world implementation. Rather than providing purely inspiration that might be difficult to execute, this constrained approach helps users bridge the gap between digital possibilities and physical creation, making physical creation more accessible and fulfilling for novice creators.

6.2 User Agency and Control in Human-AI Co-Creation

Our study shows that giving users greater control over the creative process (e.g., allowing them to select, adjust, and refine AI-generated suggestions) significantly increases their engagement and sense of authorship compared to a more passive experience. Participants reported that features like sketching, manipulating visual elements, and iterative refinement enabled them to more precisely articulate their creative intentions, leading to higher satisfaction and a stronger sense of ownership over the final outcome. In contrast, when working with a more "black-box" ChatGPT user interface, participants often felt less agency and tended to accept results that were merely "looked good", rather than actively shaping the outcome. This observation is supported by several participants' feedback, as described in Section 5.3.2. While our study did not systematically explore the level of AI assistance or autonomy, these findings indicate that preserving user agency and active participation is important for fostering productive human-AI co-creation. This aligns with previous research [57], indicating that creative tools should support and augment human creative capabilities, not attempt to replace them.

6.3 Implications for Designing Creativity Support Tools

Our user studies yielded several implications for designing future creativity support systems that integrate generative AI. First, structured exploration is essential for reducing cognitive load. The participants reported that starting from a blank canvas or prompt was challenging, whereas our step-by-step workflow, which guides users to explore associations, experiment with visual effects, arrange layouts, and refine sketches, helped externalize their thinking and provided a clear entry point for ideation. Second, balancing system suggestions with user agency is critical. The participants valued having control over the process, including the ability to select associations, adjust layouts, and add their own sketches, rather than passively receiving generated outputs. This controllability not only increased their satisfaction but also enhanced their sense of authorship. Finally, iterative co-creation encourages deeper engagement and more diverse outcomes. The participants frequently built on system-generated results, refined them, and explored variations, which led to richer and more playful creative directions. Future systems should facilitate easy revision and extension of previous outputs, supporting creativity as an ongoing conversation between the user and the AI.

6.4 Limitations and Future Work

TABLE 3
Average processing time of each module in *GenFODrawing*, calculated over five consecutive runs.

Module	Average Time (s)
Association Generation	33.98
Visual Description Generation	24.73
Layout Generation	10.72
Prompt Generation	9.29
Text-to-Image Generation	5.24
Image-to-Image Generation	0.08

While our study demonstrates promising results in human-AI co-creation, it still has several limitations. To ensure a responsive user experience, we deliberately prioritized efficiency by using Stable Diffusion v1.5 rather than employing more advanced image generation models. This choice significantly reduced image generation latency, though at the cost of some image quality and detail. In the future, we can explore hybrid strategies to balance the generation time and quality, such as generating fast, low-resolution images during early exploration and selectively invoking higher-quality models for final rendering. Our profiling further revealed that VLM and LLM inference are the primary bottlenecks causing overall latency. To provide a clearer view of where runtime costs occur, we report the average processing time of each module in Table 3, which can guide future optimization efforts. The current system primarily focuses on 2D visual features, such as shapes, and relies on the quality of object segmentation for downstream processing. As a result, our approach is effective for found objects with clear boundaries and simple outlines, including most items and some shadows, as shown in our study. However, more complex or ambiguous physical objects, such as those with overlapping elements, irregular shapes, or unclear boundaries, remain a challenge. Addressing these limitations will require integrating more advanced computer vision and generation techniques to enhance the system's comprehension of diverse physical objects and spatial relationships. While our system supports personalization through themes, composition levels, moods, and styles, and offers interaction mechanisms such as guided exploration, editable interfaces, and fine-grained sketch- and box-conditioned control, future work should explore more natural ways for AI to understand and respond to users' creative processes, leading to more intuitive and seamless human–AI co-creation. It is also important to acknowledge that, despite the positive findings, GenAI-assisted creativity support tools may introduce new challenges. Prior works [18], [40] suggest that over-reliance on AI-generated outputs can sometimes reduce user originality or inadvertently encourage convergence on similar ideas. Furthermore, some users may experience diminished agency or engagement if the system does not provide sufficient opportunities for active participation and self-expression. As such, future research should continue to examine these trade-offs and develop design strategies that both harness the strengths of generative AI and safeguard human creativity, autonomy, and diversity of thought. Overall, our findings highlight the potential of generative AI to empower human creativity. Future research can build upon these insights to explore broader creative tasks and domains, developing human–AI collaborative systems that support creative expression across various skill levels.

7 Conclusion

We have presented *GenFODrawing*, a tool designed to enhance creativity in found object drawing by incorporating everyday objects into imaginative works. We began by identifying the challenges that many individuals encounter in creating this art form, particularly the difficulties in generating creative ideas and finding appropriate reference images to bring their concepts to life on paper. Drawing from insights gathered in the formative study, our system uses AI-driven textual and visual inspirations to encourage users to explore a diverse array of ideas and visuals that resonate with their creative goals. In the comparative study, the participants utilizing *GenFODrawing* reported a more fulfilling experience in idea exploration, higher agency and engagement, and significantly enhanced creativity support and control compared to the baseline condition, where they used their own preferred tools and methods. Furthermore, the open-ended study involving twelve participants reaffirmed the system's usability, with all participants expressing that the creative process was both engaging and enjoyable. These results underscore the potential of *GenFODrawing* to support creativity and bridge the gap between physical and digital creation.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback and the user study participants for their time. This work was partially supported by grants from NSFC (62472287), Guangdong Basic and Applied Basic Research Foundation (2023A1515011297), Shenzhen Natural Science Foundation (JCYJ20250604181519025), and the Research Grants Council of the Hong Kong Special Administrative Region, China (HKUST PDFS2223-1S10).

References

[1] M. Iversen, "Readymade, found object, photograph," *Art Journal*, vol. 63, no. 2, pp. 44–57, 2004.

[2] P. M. Camic, "From trashed to treasured: A grounded theory analysis of the found object." *Psychology of Aesthetics, Creativity, and the Arts*, vol. 4, no. 2, p. 81, 2010.

[3] J. Brooker, "Found objects in art therapy," *International Journal of Art Therapy*, vol. 15, no. 1, pp. 25–35, 2010.

[4] M. Samaniego, N. Usca, J. Salguero, and W. Quevedo, "Creative thinking in art and design education: A systematic review," *Education Sciences*, vol. 14, no. 2, p. 192, 2024.

[5] M. Bat Or and O. Megides, "Found object/readymade art in the treatment of trauma and loss," *Journal of Clinical Art Therapy*, vol. 3, no. 1, p. 3, 2016.

[6] Y. J. Lee, C. L. Zitnick, and M. F. Cohen, "Shadowdraw: real-time user guidance for freehand drawing," *TOG*, vol. 30, no. 4, pp. 1–10, 2011.

[7] F. Ibarrola, T. Lawton, and K. Grace, "A collaborative, interactive and context-aware drawing agent for co-creative design," *TVCG*, 2023.

[8] J. Choi, H. Cho, J. Song, and S. M. Yoon, "Sketchhelper: Real-time stroke guidance for freehand sketch retrieval," *TMM*, vol. 21, no. 8, pp. 2083–2092, 2019.

[9] J. Xie, A. Hertzmann, W. Li, and H. Winnemöller, "Portraitsketch: Face sketching assistance for novices," in *UIST*, 2014, pp. 407–417.

[10] L. Benedetti, H. Winnemöller, M. Corsini, and R. Scopigno, "Painting with bob: assisted creativity for novices," in *UIST*, 2014, pp. 419–428.

[11] Y. Chen, K. C. Kwan, and H. Fu, "Autocompletion of repetitive stroking with image guidance," *Computational Visual Media*, vol. 9, no. 3, pp. 581–596, 2023.

[12] M. A. Mozaffari, X. Zhang, J. Cheng, and J. L. Guo, "Ganspiration: Balancing targeted and serendipitous inspiration in user interface design with style-based generative adversarial network," in *CHI*, 2022, pp. 1–15.

[13] Q. Wan and Z. Lu, "Gancollage: A gan-driven digital mood board to facilitate ideation in creativity support," in *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, 2023, pp. 136–146.

[14] D. Choi, S. Hong, J. Park, J. J. Y. Chung, and J. Kim, "Creativeconnect: Supporting reference recombination for graphic design ideation with generative ai," in *CHI*, 2024, pp. 1–25.

[15] R. E. Beaty and Y. N. Kenett, "Associative thinking at the core of creativity," *Trends in cognitive sciences*, vol. 27, no. 7, pp. 671–683, 2023.

[16] M. Benedek, T. Könen, and A. C. Neubauer, "Associative abilities underlying creativity." *Psychology of Aesthetics, Creativity, and the Arts*, vol. 6, no. 3, p. 273, 2012.

[17] H.-K. Ko, G. Park, H. Jeon, J. Jo, J. Kim, and J. Seo, "Large-scale text-to-image generation models for visual artists' creative works," in *IUI*, 2023, pp. 919–933.

[18] H. Kumar, J. Vincentius, E. Jordan, and A. Anderson, "Human creativity in the age of llms: Randomized experiments on divergent and convergent thinking," in *CHI*, 2025, pp. 1–18.

[19] J. C. Kaufman and R. J. Sternberg, *The Cambridge handbook of creativity*. Cambridge University Press, 2010.

[20] M. A. Runco, *Creativity: Theories and Themes: Research, Development, and Practice*. Elsevier, 2014.

[21] R. K. Sawyer and D. Henriksen, *Explaining creativity: The science of human innovation*. Oxford university press, 2024.

[22] B. Shneiderman, "Creativity support tools: accelerating discovery and innovation," *Communications of the ACM*, vol. 50, no. 12, pp. 20–32, 2007.

[23] D. G. Jansson and S. M. Smith, "Design fixation," *Design studies*, vol. 12, no. 1, pp. 3–11, 1991.

[24] A. Cai, S. R. Rick, J. L. Heyman, Y. Zhang, A. Filipowicz, M. Hong, M. Klenk, and T. Malone, "Designaid: Using generative ai and semantic diversity for design inspiration," in *Proceedings of The ACM Collective Intelligence Conference*, 2023, pp. 1–11.

[25] K. Son, D. Choi, T. S. Kim, Y.-H. Kim, and J. Kim, "Genquery: Supporting expressive visual search with generative models," in *CHI*, 2024, pp. 1–19.

[26] Y. Kang, Z. Sun, S. Wang, Z. Huang, Z. Wu, and X. Ma, "Metamap: Supporting visual metaphor ideation through multi-dimensional example-based exploration," in *CHI*, 2021, pp. 1–15.

[27] X. Peng, J. Koch, and W. E. Mackay, "Designprompt: Using multimodal interaction for design exploration with generative ai," in *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, 2024, pp. 804–818.

[28] J. C. Licklider, "Man-computer symbiosis," *IRE transactions on human factors in electronics*, no. 1, pp. 4–11, 1960.

[29] R. Louie, A. Coenen, C. Z. Huang, M. Terry, and C. J. Cai, "Novice-ai music co-creation via ai-steering tools for deep generative models," in *CHI*, 2020, pp. 1–13.

[30] J. Xing, M. Xia, Y. Liu, Y. Zhang, Y. Zhang, Y. He, H. Liu, H. Chen, X. Cun, X. Wang *et al.*, "Make-your-video: Customized video generation using textual and structural guidance," *TVCG*, 2024.

[31] S. Wang, S. Menon, T. Long, K. Henderson, D. Li, K. Crowston, M. Hansen, J. V. Nickerson, and L. B. Chilton, "Reelframer: Human-ai co-creation for news-to-video translation," in *CHI*, 2024, pp. 1–20.

[32] A. Yuan, A. Coenen, E. Reif, and D. Ippolito, "Wordcraft: story writing with large language models," in *IUI*, 2022, pp. 841–852.

[33] Z. Zhang, J. Gao, R. S. Dhaliwal, and T. J.-J. Li, "Visar: A human-ai argumentative writing assistant with visual programming and rapid draft prototyping," in *UIST*, 2023, pp. 1–30.

[34] Y. Liu, Z. Wen, L. Weng, O. Woodman, Y. Yang, and W. Chen, "Sprout: an interactive authoring tool for generating programming tutorials with the visualization of large language models," *TVCG*, 2024.

[35] L. Cheng, D. Deng, X. Xie, R. Qiu, M. Xu, and Y. Wu, "Snil: Generating sports news from insights with large language models," *TVCG*, 2024.

[36] R. Huang, H. Lin, C. Chen, K. Zhang, and W. Zeng, "Plantography: Incorporating iterative design process into generative artificial intelligence for landscape rendering," in *CHI*, 2024, pp. 1–19.

[37] S. Xiao, L. Wang, X. Ma, and W. Zeng, "Typedance: Creating semantic typographic logos from image through personalized generation," in *CHI*, 2024, pp. 1–18.

[38] J. Liao, P. Hansen, and C. Chai, "A framework of artificial intelligence augmented design support," *Human–Computer Interaction*, vol. 35, no. 5-6, pp. 511–544, 2020.

[39] J. Kim and M. L. Maher, "The effect of ai-based inspiration on human design ideation," *International Journal of Design Creativity and Innovation*, vol. 11, no. 2, pp. 81–98, 2023.

[40] A. R. Doshi and O. P. Hauser, "Generative ai enhances individual creativity but reduces the collective diversity of novel content," *Science Advances*, vol. 10, no. 28, p. eadn5290, 2024.

[41] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *NeurIPS*, vol. 35, pp. 24 824–24 837, 2022.

[42] T. B. Brown, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[43] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," *arXiv preprint arXiv:2104.08691*, 2021.

[44] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *ICCV*, 2023, pp. 3836–3847.

[45] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *AAAI*, vol. 38, no. 5, 2024, pp. 4296–4304.

[46] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "Gligen: Open-set grounded text-to-image generation," in *CVPR*, 2023, pp. 22 511–22 521.

[47] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *ICCV*, 2023, pp. 4015–4026.

[48] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[49] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language-image learning," in *CVPR*, 2023, pp. 2818–2829.

[50] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.

[51] V. Sarukkai, L. Yuan, M. Tang, M. Agrawala, and K. Fatahalian, "Block and detail: Scaffolding sketch-to-image generation," in *UIST*, 2024, pp. 1–13.

[52] Q. Phung, S. Ge, and J.-B. Huang, "Grounded text-to-image synthesis with attention refocusing," in *CVPR*, 2024, pp. 7932–7942.

[53] awacke1, "Image-to-line-drawings," hugging face spaces, 2023. [Online]. Available: https://huggingface.co/spaces/awacke1/Image-to-Line-Drawings

[54] E. Cherry and C. Latulipe, "Quantifying the creativity support of digital tools through the creativity support index," *TOCHI*, vol. 21, no. 4, pp. 1–25, 2014.

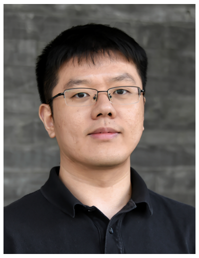
[55] S. Hart, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," *Human mental workload/Elsevier*, 1988.

[56] J. Brooke *et al.*, "Sus-a quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

[57] F. Gmeiner, H. Yang, L. Yao, K. Holstein, and N. Martelaro, "Exploring challenges and opportunities to support designers in learning to co-create with ai-based manufacturing design tools," in *CHI*, 2023, pp. 1–20.

Jiaye Leng is a PhD candidate at the School of Creative Media, City University of Hong Kong. Before joining CityUHK, he received a Master's Degree in Computer Technology from Beihang University and a Bachelor's Degree in Electronic Information Science and Technology from China University of Mining and Technology. His research interests are Human-AI Interaction and Creativity Support.

Hui Ye is an Assistant Professor at the Department of Interactive Media, School of Communication, Hong Kong Baptist University. She was previously an RGC Postdoctoral Fellow at the Division of Arts and Machine Creativity of Hong Kong University of Science and Technology and the School of Creative Media of City University of Hong Kong (CityUHK). She obtained her PhD degree from CityUHK and a Bachelor's degree from the University of Science and Technology of China. Her research interests lie in the intersection of Human-Computer Interaction and Computer Graphics. She has published papers in ACM SIGCHI, ACM TOG, IEEE TVCG, etc.



Pengfei Xu is an Associate Professor at the College of Computer Science and Software Engineering, Shenzhen University. He received his Bachelor's degree in Math from Zhejiang University, China, in 2009 and his Ph.D. in Computer Science from the Hong Kong University of Science and Technology in 2015. His primary research lies in Human-Computer Interaction and Computer Graphics.

Miu-Ling Lam is an Associate Professor in the School of Creative Media at City University of Hong Kong. She holds a PhD degree in Automation and Computer-Aided Engineering from The Chinese University of Hong Kong. She was formerly a Croucher postdoctoral research fellow at the University of California, Los Angeles. Her research interests include computational imaging, light field technology, and robotics.

Hongbo Fu is a Professor at the Division of Arts and Machine Creativity, Hong Kong University of Science and Technology. Before joining HKUST, he worked at the School of Creative Media, City University of Hong Kong, for over 15 years. He had postdoctoral research training at the Imager Lab, University of British Columbia, Canada, and the Department of Computer Graphics, Max-Planck-Institut Informatik, Germany. He received a Ph.D. degree in computer science from the Hong Kong University of Science and Technology in 2007 and a BS degree in information sciences from Peking University, China, in 2002. His primary research interests fall in computer graphics, human-computer interaction, and computer vision. He has served as an associate editor of The Visual Computer, Computers & Graphics, and Computer Graphics Forum.