ContextCam: Bridging Context Awareness with Creative Human-AI Image Co-Creation

Xianzhe Fan
Tsinghua University
Beijing, China
fxz21@mails.tsinghua.edu.cn

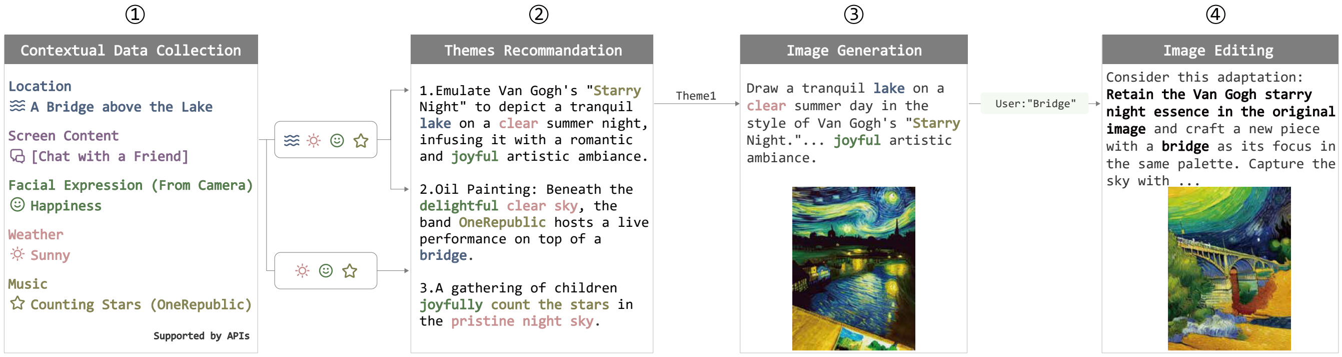
Zihan Wu
University of Michigan
Ann Arbor, Michigan, United States
ziwu@umich.edu

Chun Yu
Tsinghua University
Beijing, China
chunyu@tsinghua.edu.cn

Fenggui Rao
China Academy Of Art
HangZhou, ZheJiang, China
admin@whitesir.cn

Weinan Shi*
Tsinghua University
Beijing, China
swn@tsinghua.edu.cn

Teng Tu
Tsinghua University
Beijing, China
leotuteng@126.com



1.Emulate Van Gogh's "Starry Night" to depict a tranquil lake on a clear summer night, infusing it with a romantic and joyful artistic ambiance.

2.Oil Painting: Beneath the delightful clear sky, the band OneRepublic hosts a live performance on top of a bridge.

3.A gathering of children joyfully count the stars in the pristine night sky.

Draw a tranquil lake on a clear summer day in the style of Van Gogh's "Starry Night."... joyful artistic ambiance.

Consider this adaptation: **Retain the Van Gogh starry night essence in the original image** and craft a new piece with a **bridge** as its focus in the same palette. Capture the sky with ...

Figure 1: On a *sunny* night, Alex stands on *a bridge above the lake*, *texting their friend* and *enjoying the song "Counting Stars" by OneRepublic*, *smiling*. They want to capture the beautiful moment, then open the ContextCam, and collaborate with it to create a piece of art inspired by their current context. During the "framing" phase, ContextCam extracts relevant contextual data and proposes three themes. Alex selects theme one, asks ContextCam to create the image, and polishes it through in-depth discussion with ContextCam.

ABSTRACT

The rapid advancement of AI-generated content (AIGC) promises to transform various aspects of human life significantly. This work particularly focuses on the potential of AIGC to revolutionize image creation, such as photography and self-expression. We introduce ContextCam, a novel human-AI image co-creation system that integrates context awareness with mainstream AIGC technologies like Stable Diffusion. ContextCam provides user's image creation process with inspiration by extracting relevant contextual data, and leverages Large Language Model-based (LLM) multi-agents to co-create images with the user. A study with 16 participants and 136 scenarios revealed that ContextCam was well-received, showcasing personalized and diverse outputs as well as interesting user behavior patterns. Participants provided positive feedback on their engagement and enjoyment when using ContextCam, and acknowledged its ability to inspire creativity.

CCS CONCEPTS

• **Human-centered computing** → *Ubiquitous and mobile computing systems and tools.*

KEYWORDS

Human-AI Co-Creation, Context-Aware Systems, Image Generation and Editing, LLM-Based Multi-Agent Systems

ACM Reference Format:
Xianzhe Fan, Zihan Wu, Chun Yu, Fenggui Rao, Weinan Shi, and Teng Tu. 2024. ContextCam: Bridging Context Awareness with Creative Human-AI Image Co-Creation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3613904.3642129

1 INTRODUCTION

Images have become deeply integrated into our lives. Whether through painting, photography, or digital technology, the creation

---

[*] Corresponding Author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI '24, May 11–16, 2024, Honolulu, HI, USA
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642129

and generation of images often aim to convey certain information, emotions, or affections. The creation can stem from personal visual experiences, feelings, or imagination of the creator. Certain environments and events can inspire us to create our own images. However, not all of us are capable of turning the transient inspiration we feel from our surroundings into a piece of art.

Thankfully, the advent of AI-generated content (AIGC) has brought about profound changes in the way we create images [7]. Creating an image has become a task that people can complete without much effort, particularly using image generation models like Stable Diffusion [48, 51] and image editing models like ControlNet [71].

To help people capture the inspiration from their environment, and create images inspired by the place they are at, the emotions they feel, the music they hear and so on, we pose an essential question: Can context awareness, combined with human-AI co-creation, aid in creating more personalized and engaging images?

To gather users' insights on embedding context awareness in human-AI image co-creation, we first conducted a formative study on 23 participants. We collected their motivations for AI image generation, needs, and opinions on existing image generators. We also adopted participatory design strategies, asking participants to envision an ideal context-aware human-AI image co-creation system, installed on their mobile phones, that would be capable of creating images inspired by personal contexts.

Based on the users' feedback and ideas from the formative study, we introduce **ContextCam**, **a human-AI co-creation system that incorporates context awareness to generate artistic images**. ContextCam transcends traditional image generation by integrating environmental information, such as location and weather, with personal states, including facial expression, music collected by sensors, and screen contents on the phone. The co-creation system offers a canvas where human creativity meets AI innovation. This mutual inspiration and learning between the user and AI enhance the creation's depth and personalization.

The workflow of ContextCam contains two phases: **framing** and **focusing**. In the "framing" phase, ContextCam deduces user intent based on in-situ contextual data and user directives, then generates three themes for the image as inspirations for the user. In the "focusing" phase, users collaborate with AI to create images that meet their satisfaction. In this co-creation process, users can request edits to the image or seek AI's ideas effortlessly through natural language commands and straightforward selections.

To evaluate the performance of our system in practical settings, and understand user behavior patterns within it, we conducted a study involving 16 participants and 136 scenarios. Our findings showed high user satisfaction with the images produced by ContextCam. In 92.9% of the scenarios, participants picked ContextCam's topic recommendations. The average user input was 1.1 words per interaction. Participants also rated high overall enjoyment, engagement, usability, and inspiration.

By analyzing interaction log data and conducting interviews, we have gained insights into how individuals employ and perceive contextual data in their collaborative creative processes with AI. Our research highlights the role that contextual information plays in impacting image themes, influencing user behaviors, acting as a source of creative inspiration, and enriching collaborative experiences between humans and AI. This exploration aims to shed light on the profound impact of context awareness on inspiration and user engagement in human-AI image co-creation.

Our work's contributions are as follows:

- We propose a novel paradigm that integrates context awareness with human-AI collaborative image creation, based on a formative study (N=23) aimed at exploring user insights on human-AI image co-creation and the limitations of existing AI image generators. Building on the formative study, we have summarized six design guidelines for our paradigm.
- We adopt an LLM-based multi-agent approach to help users create images by providing image ideations and generating text-to-image/image-to-image prompts in ContextCam, our human-AI image co-creation system with context awareness.
- Our user study (N=16) in a real-world setting validates ContextCam's effectiveness in providing satisfactory image-generation results, and presenting personalized and diverse image outputs co-created by humans and AI. We provide insights on users' behavior patterns demonstrated while using ContextCam, as well as how context awareness affects human-AI collaboration and creative process.

The paper unfolds as follows: Section 2 reviews relevant literature, highlighting the growing field of human-AI co-creation, context-aware systems, image generation and editing, and LLM-based multi-agent systems. Section 3 details our findings from the formative study that informed the design of ContextCam. Then, we introduce the system design of ContextCam in section 4. Section 5 presents our comprehensive user study, followed by a detailed result presentation and discussion in section 6. This exploration concludes with Sections 7 and 8, shedding light on the potential of context-aware human-AI image co-creation.

2 RELATED WORKS

Our multi-agent system integrates context awareness into the process of human-AI co-creation, and enables the generation of AI-created images that resonate with the user's emotions and experiences, thereby enhancing creativity and personalization. We summarize related work into four parts: (1) human-AI co-creation, (2) context-aware systems, (3) image generation and editing, and (4) LLM-based multi-agent systems.

2.1 Human-AI Co-Creation

Advancements in technology have fostered human-AI co-creation across various domains. Compared to solely AI-driven solutions, this collaboration offers users a greater sense of involvement and agency. Prior works have demonstrated this synergy in enhancing human creativity in several areas, such as refining drawings [22, 25], enriching writings [11, 17, 18, 44, 57], and improving designs [10, 21]. In these examples, humans and AI fuel each other's creativity, resulting in outcomes that satisfies the users [27, 69].

The introduction of LLMs and image generation models has further revolutionized the domain of human-AI co-creation [67]. One way to assist the co-creation process is to use LLMs to provide suggestions to generate the text-to-image prompts. An example is Opal [33], a generative workflow for news illustration, demonstrating how LLMs can provide prompt exploration support for text-to-image generations. Similarly, 3DALL-E [34] generates LLM-provided image inspiration for CAD and product design by helping users craft text prompts with design language and image prompts connected to their work in progress. While these tools [24, 32–34, 55] successfully leverage LLMs to help users craft ideal prompts and improve the co-creation process, they fail to take advantage of contextual information to optimize the co-creation experience.

Existing research suggests that incorporating contextual data into human-AI collaboration can spark greater inspiration [8, 26]. However, this aspect remains underexplored using current generative models like LLMs and diffusion models. Our research introduces contextual information into the human-AI co-creation process, crafting AI-generated images that echo users' inspirations from their physical environments and emotions. By doing so, our work fills a void in the existing literature and proposes a novel direction for more adaptive and context-aware AI tools in co-creation domains.

2.2 Context-Aware Systems

Context-aware systems leverage user context to offer personalized services [14]. These systems gather diverse information through different sensors and APIs, such as cameras for facial expression, microphones for voice or music recognition, and GPS for location data. Such rich information, combined with algorithms designed to understand user intentions [9], enables relatively accurate prediction of user needs. Given their success in tailoring user experiences, context-aware systems have widespread application across various domains [1, 6, 12, 29], from mobile assistants offering real-time, context-sensitive recommendations [1] to smart home automation [35].

Context awareness has great potential in image generation, which can improve user engagement and refine their interactive experience [41, 68]. For example, UbiFit Garden [13] and WhoIsZuki [40] uses on-body sensing to detect users' daily activities, providing pertinent visual feedback to encourage increased physical activity. Similarly, MateBot [61] harnesses data from smartphone sensors, improving the interaction quality between users and virtual robots through anthropomorphic images based on real-time context. Context awareness can boost the creative journey and foster innovation [56]. Through evaluating several leading creativity support tools, Sielis et al. discovered that the inclusion of context awareness significantly boosts real-time interaction, enabling tools to proactively adapt to user needs. We extend this concept into AIGC, an area still in its infancy concerning context-aware applications. As of now, we were unable to locate similar work in the domain of human-AI co-creation that combines mainstream image generation tools or models, such as Stable Diffusion [51] and Midjourney [38], with context-aware systems. Our context-aware system integrates diffusion models' powerful image generation capabilities and LLM's prowess in creative inspiration and user intent comprehension, aiming to create a personalized and engaging human-AI image co-creation experience.

2.3 Image Generation and Editing

The advances in image generation and editing tasks have the potential to revolutionize the way we create and manipulate images. Text-to-image diffusion models [48, 50, 51] achieve state-of-the-art image generation results by encoding text inputs into latent vectors via pre-trained language models like CLIP [49]. Image editing models such as ControlNet [71] and T2I-Adapter [39] add additional conditions to control the outcome of diffusion models, which allow users to edit images based on textual input. Prior research like Opal [33], GenAssist [24], and 3DALL-E [34] use LLMs to generate the prompts for users in the image creation process, resulting in high-quality outputs. Building on this, our system integrates contextual information into the workflow of LLM-based prompts generation for human-AI image co-creation, inspiring more creativity in the process.

2.4 Large Language Model-based (LLM) Multi-Agent Systems

Large Language Models (LLMs) refer to Transformer language models that contain hundreds of billions (or more) of parameters, which are trained on massive text data [53], such as ChatGPT [42] and LaMDA [59]. With the instruction-following capabilities provided by in-context learning [15] and the logical reasoning abilities offered by Chain of Thought [62], LLMs have set new benchmarks in natural language understanding and generation.

A Multi-Agent System is an extension of the agent technology where a group of connected autonomous agents act in an environment to achieve a common goal [46]. Current research and trends suggest a promising future for collaborative interactions between humans and LLM-based multi-agent systems [23, 31, 58, 66]. These systems [4, 67] combine the strengths of LLMs in language generation, logical reasoning, and creativity with human cognition and insight, efficiently addressing multi-functional and multi-stage tasks. Wu et al. [67] introduce the concept of chaining LLM steps together, where the output of one step becomes the input for the next, thus aggregating the gains per step. In this interactive multi-agent system, users can modify chains, along with their intermediate results, in a modular way. Building on Wu et al.'s research, Baek et al. [4] presented a mixed-initiative system that allows step-by-step crafting of text-to-image prompts. We aim to adopt a similar multi-agent approach that can lead the process if the user desires and provide image ideations upon request, as well as craft prompts for high-quality image generation and editing.

Moreover, the fusion of multi-agent dynamics with context-aware computing allows for a more intuitive and efficient human-AI collaboration, particularly in complex tasks [16]. Building on the prior research, we specifically introduce context awareness and have designed and implemented a multi-agent system capable of co-creating images with humans. Based on the user's natural language commands and contextual data, the system engages in user discussions, yielding image production as an output. Our work combines the strengths of LLMs, vision models, and humans, aiming to create desirable image outputs.

3 FORMATIVE STUDY

We first conducted a formative study to gather design insights for a context-aware human-AI co-creation system from participants. We hope to understand 1) the prevailing motivations and needs for AI image generation, 2) the limitations of current AI image generation, and 3) participants' vision for a context-aware human-AI image co-creation system.

3.1 Participants and Procedure

We recruited participants by sending recruitment messages in student group chats we accessed, and encouraged students to repost these messages on social media platforms. A total of 23 participants (F1 - F23, 12 male participants, and 11 female participants), aged between 19 and 46 (SD=8.80), signed up to be included in the formative study[1]. All participants had experience using AI image-generation tools. We rated their experience level as beginner (N=8, used AI image generator no more than 3 times), intermediate (N=7, used more than 3 times and are familiar with its basic functionalities), and expert (N=8, used AI image generators frequently and have posted AI-generated artwork online that received positive reviews). All research procedures were ethically approved, and we obtained participants' consent prior to the study. All participants received compensation for their time.

We conducted a semi-structured interview with each participant. Each interview session lasted approximately 30 minutes through a video conferencing software. The process resulted in 12 hours of audio recordings, complemented by detailed observation notes. During each interview, we focused on four key themes: 1) participants' motivations and needs related to AI image generation, 2) their perspectives on the current limitations of AI image generation, 3) imaginations of co-creating images with AI in different scenarios, and 4) expectations for experience design of human-AI image co-creation system. Our interview guide included a mix of open-ended and closed questions, allowing us to delve into specific topics while maintaining flexibility.

3.2 Data Analysis and Results

3.2.1 Motivations, Needs, and Limitations of Current AI Image Generation. We investigated participants' motives for using AI image generators, method purposes employed when using AI image generators, preferred subjects, and preferred styles. One researcher conducted open coding [28] of the interview transcriptions to generate insights. We also used the affinity diagram to organize our findings [37]. We identified recurring keywords and themes including motives, method purposes, preferred subjects, and preferred styles. These identified elements were then categorized and assigned representative codes, like "Motives - Curiosity" or "Method Purposes - Illustrating writings". The result is shown in Table 1.

Table 1: Exploration of participants' motives, method purposes, and preferred subjects and styles of AI image generation.

Theme	Response (Frequency)
Motives	Curiosity (16/23), Entertainment (13/23), For work purposes (6/23)
Method Purposes	Illustrating writings (e.g., poems and articles) (7/23), Creating artworks for personal expression and aesthetic exploration (11/23), Designing (e.g., logo design, industrial design, and poster design) (7/23)
Preferred Subjects	Landscapes (8/23), Beauties (9/23), Portraits (10/23), Sci-fi (14/23), Surrealism (6/23)
Preferred Styles	Anime (16/23), Oil painting (10/23), Realism (12/23), Chinese traditional style (6/23)

Participants expressed concerns about the trade-off between image quality and interaction complexity in AI image generation. Fifteen participants found crafting the proper input prompts challenging and often had trouble creating desired images. F23 commented, "While platforms like stable-diffusion-webui [3] are very powerful, their complicated interfaces and a large number of parameters undermine the intuitive user experience." Similarly, five participants (F1, F3, F9, F20, F21) mentioned that some natural language-based image generation and editing systems, such as "Visual ChatGPT [65]" and "HuggingGPT [54]" don't always meet their expectations in terms of output quality. Another issue is the personalization of AI-generated images. As F5 commented, "ImageBind [19] spans across six modalities, but it can't offer personalized recommendations for my specific context. I have to manually type in a lot of contextual information, which affects my user experience."

3.2.2 Imaginations and Expectations for a Context-Aware Human-AI Image Co-Creation System. We invited participants to envision a human-AI co-creation system capable of creating images inspired by personal contexts. First, we asked participants to imagine co-creating images with AI in different scenarios. After cutting off the irrelevant and redundant components of audio recordings, we finally got 33 ideas from participants, some of which are presented in Table 2. We adopted open coding [28] to classify the 33 ideas and summarize the recurring contextual information from the interview transcriptions. We identified five types of contextual information commonly mentioned: **location, facial expression, music, screen content, and weather.**

Next, we asked them about their expectations for experience design. We conducted thematic analysis [5] on the interview transcriptions to extract key themes and gain deeper insights into the participants' suggestions. Based on the findings, we summarized six design guidelines for a context-aware human-AI image co-creation system:

1. Require less user burden, such as fewer text input and click count. Integrate shortcuts for frequently used tasks and use a conversational interface to make interactions more natural and efficient. It would be preferable for the system to run on mobile devices.
2. Design a well-functioning intent understanding and image generation mechanism. Ensure the system generates high-quality images by adhering closely to the user's needs and preferences.
3. Accommodate a variety of input methods, including voice commands, textual input, and image-based inputs such as photographs and sketches, to enhance usability and convenience.
4. Offer a range of suggestions to users. Each conversation should start with several brief image topics and develop detailed suggestions based on user feedback and interaction.

---

[1] The details of participants are in the supplementary material.

Table 2: Some interesting ideas about context-aware human-AI image co-creation. We used open coding [28] to deduce the themes.

Theme	Participants' Ideas	Context
Travel	1) "I wish the AI could capture the valuable moments of my travels. Imagine documenting a trip to Bali through AI-generated images without ever taking a photo!" (F2) 2) "When I'm in the city, the AI should sense the urban vibes around me, possibly depicting the tall skyscrapers, busy streets, and neon lights. In contrast, if I'm in the countryside, I'd love for it to paint serene landscapes, vast fields, and peaceful sunsets." (F20)	**Location**
Design	"Being a designer, standing in an undecorated room, it would be revolutionary if the AI could sense the space and interpret my rough ideas, turning them into refined designs that perfectly fit that specific environment." (F17)	**Location**
Express My Feelings	"Sometimes words cannot express my feelings. I hope the AI can sense the current weather and read my facial expressions, creating an image that resonates with my emotions." (F19)	**Weather, Facial Expression**
Music	1) "I imagine the AI producing a vibrant, abstract painting while I am at a live rock concert or a serene watercolor during a classical music performance." (F10) 2) "Listening to jazz always evokes specific visuals for me. I would love to see the AI's interpretation of my favorite tracks." (F13)	**Music**
Illustrate for My Text	1) "It would be amazing if the AI could read the script or poem I am writing on my phone and then generate an accompanying artwork." (F11) 2) "Finding the image to perfectly convey my message is often challenging when I post updates on my WeChat Moment. It would be fantastic if the AI could craft an illustration tailored to my words, adding a personal touch to my shared moments." (F14)	**Screen Content**
Weather	"I hope the AI will illustrate a cozy indoor scene on rainy days and depict a bright landscape on sunny days." (F4)	**Weather**
News	"Imagine the AI scanning the headlines or articles from my phone and crafting a comic strip or illustration that represents the main event or theme. This would offer a quick visual snapshot of the day's major stories, making news consumption more engaging and efficient." (F15)	**Screen Content**
Dining Experience	"Imagine having dinner at a seaside restaurant. The AI could combine the sound of waves, my current location, and my contented expression to depict a beautiful oceanside feast scene." (F9)	**Location, Facial Expression**
Activities and Events	"When attending a local festival or event, the AI could use ambient sounds, my location, and calendar to depict vibrant celebrations, dances, and festivities scenes." (F23)	**Music, Location**

(5) Focus on the most relevant contextual information based on user intent and consider excluding less relevant context to optimize the user experience and reduce information overload.

(6) Allow users to control which contextual information is utilized by the system and apply a clear visual indication, such as dimming unused contextual icons, to show when certain data is not being employed.

4 SYSTEM DESIGN AND IMPLEMENTATION OF CONTEXTCAM

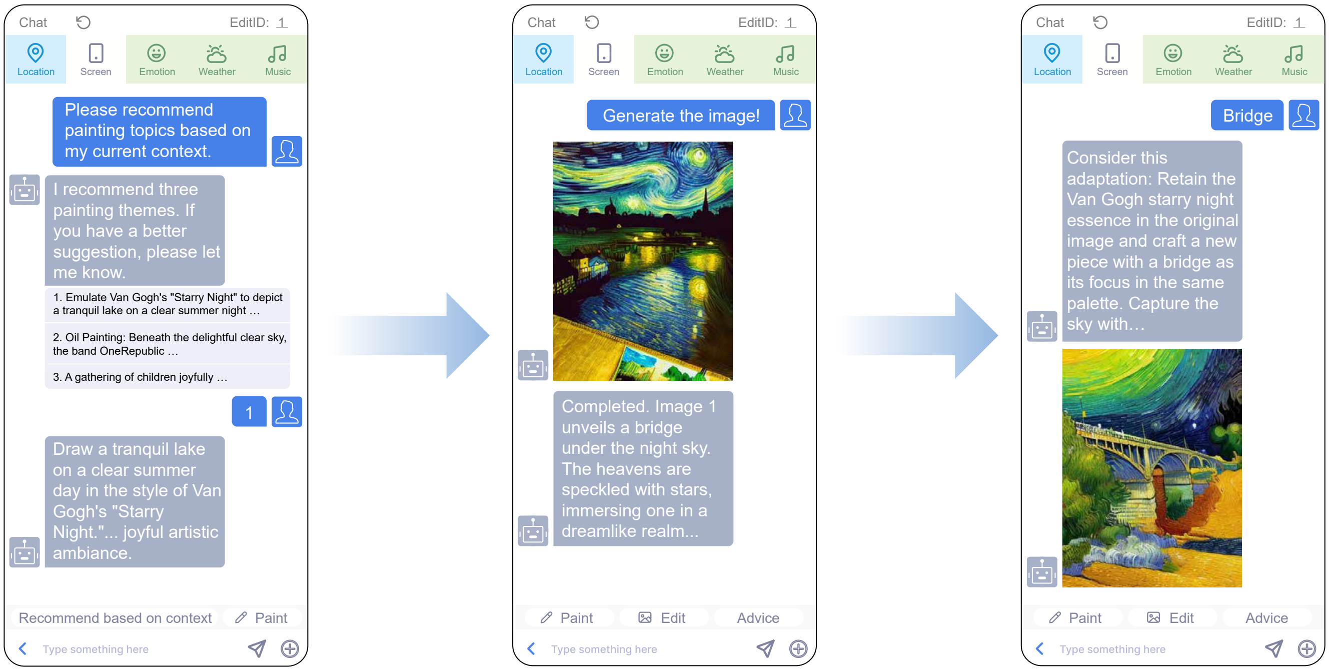
Following these guidelines, we introduce ContextCam, a context-aware multi-agent system for image co-creation between humans and AI. This section will first illustrate our system's workflow through a use case, and then detail the underlying mechanisms of each agent.

4.1 Illustrating ContextCam Through a Use Case

This section introduces how ContextCam works through a use case depicted in Figure 2.

Imagine a ***sunny night*** *(Weather)*. Alex, an engineering student with no background in art, was standing on ***a bridge above the lake*** *(Location)*, ***texting their friend*** *(Screen Content)* and enjoying ***OneRepublic's "Counting Stars,"*** *(Music Recognition)* with ***a smile*** *(Camera)*. Maybe it was the sunshine and the scenery, the beautiful music, the uplifting conversation, or just the great mood; Alex suddenly felt deeply moved by the moment and wanted to capture this feeling. To do this, Alex pressed a button on the shortcut menu (Figure 2(a)) that sent ContextCam a message: *"Please recommend painting topics based on my current context."*

Upon receiving the command, ContextCam initiated the "framing" phase. First, ContextCam retrieved contextual information that might be relevant to Alex's painting topics. The color-coded icons that indicated the status of the contexts were shown on the top of the screen: Green indicates contextual information used in all topics; Blue denotes contextual information used in Topics 1 and 2; Yellow indicates contextual information only used in Topic 3; White represents irrelevant information; Grey signifies information that the user has actively closed. In this case, ContextCam decided that Alex's location, facial expression (labeled as "emotion" in Figure 2), current weather and music were relevant.



Please recommend painting topics based on my current context.

I recommend three painting themes. If you have a better suggestion, please let me know.

1. Emulate Van Gogh's "Starry Night" to depict a tranquil lake on a clear summer night …

2. Oil Painting: Beneath the delightful clear sky, the band OneRepublic …

3. A gathering of children joyfully …

1

Draw a tranquil lake on a clear summer day in the style of Van Gogh's "Starry Night."... joyful artistic ambiance.

Generate the image!

Completed. Image 1 unveils a bridge under the night sky. The heavens are speckled with stars, immersing one in a dreamlike realm...

Bridge

Consider this adaptation: Retain the Van Gogh starry night essence in the original image and craft a new piece with a bridge as its focus in the same palette. Capture the sky with…

(a) Choose a Creation Topic

(b) Generate the Image

(c) Edit the Image

Figure 2: Screenshots of ContextCam's workflow, including assisting the user in choosing the creation topic, generating the image, and editing it based on the user's feedback. When editing the image, Alex mentioned "Bridge," choosing to focus on the "Bridge" in the initial image. For an enhanced user experience, the interface provides four intuitive quick-response buttons: "Recommend based on context" for creating images only based on the current contextual data instead of other topics; "Paint" for image generation; "Edit" for editing the image specified in "EditID"; "Advice" for seeking advice about the detailed image ideation.

With the contextual information, ContextCam recommended three themes to Alex (Figure 2(a)):

1. *Emulate Van Gogh's "Starry Night" to depict a tranquil lake on a clear summer night, infusing it with a romantic and joyful artistic ambiance.*
2. *Oil Painting: Beneath the delightful clear sky, the band OneRepublic hosts a live performance on top of a bridge.*
3. *A gathering of children joyfully count the stars in the pristine night sky.*

Alex loved the first recommendation and clicked on it, sending the number "1" to ContextCam. Then, focusing on topic 1, ContextCam entered the "focusing" phase to discuss and iterate on the image with Alex. ContextCam first proposed (Figure 2(a)): *"Draw a tranquil lake on a clear summer day in the style of Van Gogh's 'Starry Night.' ... joyful artistic ambiance."* Alex found the proposal pretty interesting and clicked the "Paint" button on the shortcut menu, automatically sending a message: *"Generate the image!"* ContextCam generated an image and replied with its description: *"Completed. Image 1 unveils a bridge under the night sky. The heavens are speckled with stars, immersing one in a dreamlike realm ..."*

Alex felt inspired and typed in *"Bridge"* for modifying the image. With this information, ContextCam proposed a description to help the user modify the image: *"Consider this adaptation: Retain the Van Gogh starry night essence in the original image and craft a new piece with a bridge as its focus in the same palette. Capture the sky with ..."* It also provided the updated image. They continued the discussion around the image to perfect it, each round Alex provided some new ideas for the image, and ContextCam kept proposing modified descriptions and updated images based on Alex's input. The two of them bounced ideas off each other for a few rounds. Finally, Alex was pleased with the resulting image and captured the precious moment. After their conversation ended, ContextCam transformed this conversation history into a summary of user preferences, so the next time Alex consults ContextCam, it could include that in the decision-making process.

4.2 The Multi-Agent System Design

In response to the complex procedures and knowledge required for image creation, we adopted a multi-agent system design based on Weng's LLM-powered system [63]. LLM serves as the brain of each agent in Weng's system, augmented by several crucial components: memory, planning, tools, and action. We have modified

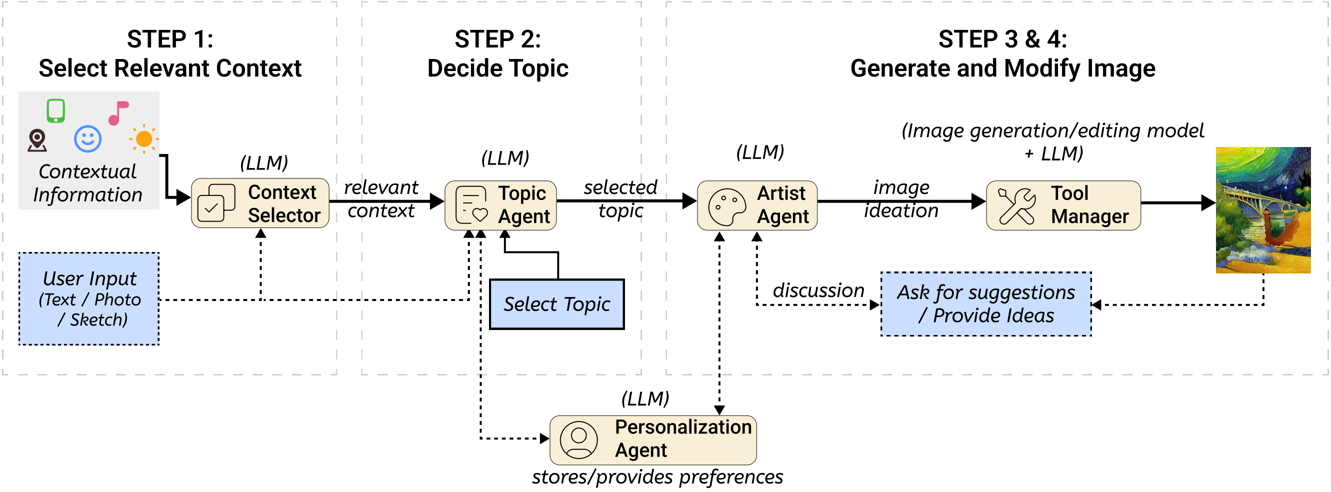


Figure 3: The multi-agent workflow of ContextCam. We roughly divided the workflow into four steps, corresponding to Figure 1. Each yellow rectangle represents an agent in ContextCam. Each blue rectangle represents user input to ContextCam: blue rectangles with dotted borders suggest the input is optional. To start the image creation process, users can either command ContextCam to start creating entirely based on their current context, or, as a supplement, they can choose to input text, a photo, or a sketch. Context Selector then extracts relevant contextual information (step 1). Then, Topic Agent provides topic candidates to the user and receives user feedback to decide on the final topic (step 2). Finally, Artist Agent creates the image ideation based on the topic, and passes it on to the tool manager to select the appropriate image-generation or image-editing model and generate a prompt for the model to generate a new image. Based on the image, the user can exchange more ideas with Artist Agent to reach the final desired image (steps 3 and 4). Personalization agent keeps track of past interactions, synthesizes user preferences, and provides it to other agents.

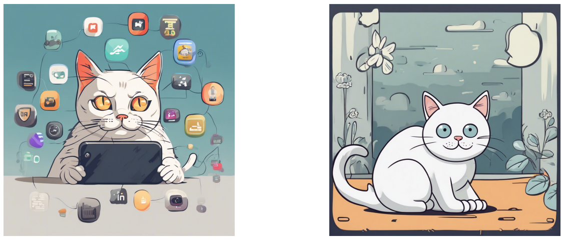
the names and functions of some components in Weng's definition of "agent". In our design, each agent is responsible for a distinct task, potentially integrating one or more of the following components:

- **Planning**: deconstructing task into manageable and actionable subtargets.
- **Memory**: accumulating and storing knowledge, whether derived internally or retrieved from external sources.
- **Tool use**: adopting appropriate external tools to complete the task.
- **Evolving**: enhancing performance and refining planning strategies from system feedback.

Figure 3 shows ContextCam's workflow. The system comprises two creation phases (**"framing": STEP 1,2 and "focusing": STEP 3,4**) and five agents (**Context Selector, Topic Agent, Tool Manager, Artist Agent, and Personalization Agent**).

4.2.1 "Framing" Phase - Context Selector. From John Dewey's "Art as Experience," we learned that image creation is deeply connected to its environment [47]. Our system, ContextCam, exemplifies this idea by being context-aware and drawing inspiration from its surroundings. ContextCam combines what users want with where they are, what they are listening to, how they feel, and what they see.

However, as shown in Figure 4, not all contexts are relevant to the subject of the user's drawing. We introduce the **Context Selector** to prevent such issues. Context Selector leverages LLM to select contextual information relevant to the current art creation process. We adopted the Few-Shot-CoT [72] approach to improve LLM's ability to solve complex problems, where LLM performs Chain-of-Thought [62] reasoning with several demonstrations. Based on the results of our formative study, we defined a series of user intention modes, each accompanied by a "standard vector". The output vector type is chosen automatically via a prompt to the



(a) Without Context Selector: "A cat immersed in the mobile app market."

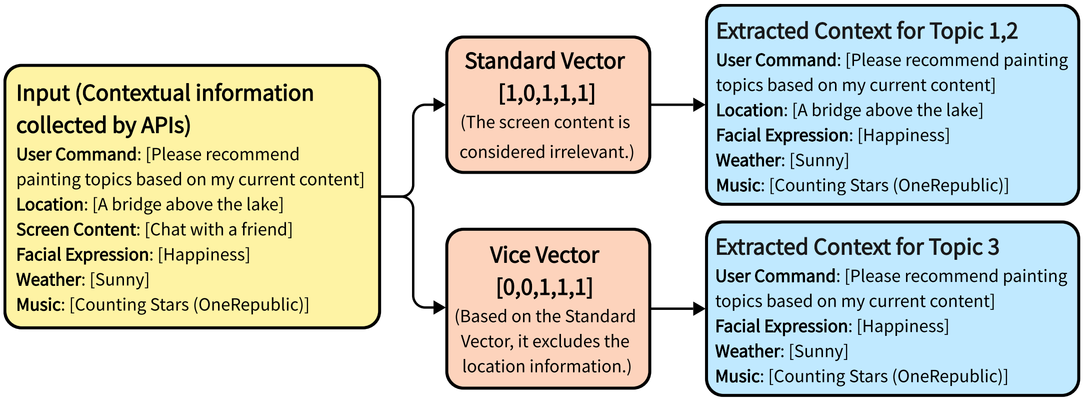
(b) With Context Selector: "A cute cat in the garden."

Figure 4: When a user intends to create an image themed around cats, they activate ContextCam's float window on their mobile home screen. However, ContextCam might inadvertently capture irrelevant text, such as "mobile app market". If ContextCam does not filter out the irrelevant information, LLM will be confused, leading to inaccurate results like (a). Such recommendations tend to be off-putting for users.

Table 3: System prompt for Context Selector.

Given user information across five modalities: *location, screen content, facial expression, weather, and music*, we have eight predefined scenarios. Each scenario is accompanied by a 5-dimensional vector, where each element is either 0 (irrelevant for the painting) or 1 (relevant). *Let's think step by step.* Determine the appropriate painting scenario by analyzing the user's context and selecting or devising a corresponding 5-dimensional vector. If none of the eight scenarios fits, use your judgment to create a relevant vector based on the *user command*.

User Intention Mode	**Description**	**Standard Vector**
Free Creation Mode	The user requests themes to be recommended based entirely on the current context.	[1,1,1,1,1]
Art Mode	The *user command* contains professional art vocabulary, and the *location* is residential, schools, galleries, or other life and art places.	[0,0,1,1,1]
Textual Mode	The *screen content* includes articles, poetry, or speeches, and the user requests illustrations for the texts on screen. The *location* could be residential, office buildings, schools, coffee shops, or other life and office places.	[0,1,0,0,0]
Architect Mode	The *user command* is related to architectural or environmental design, and the *location* is outdoors (near buildings or parks).	[1,0,0,1,0]
Travel Mode	The *location* is outdoors, such as in parks, and the *user command* involves drawing immediate surroundings or current locations, not distant or unrelated places.	[1,0,1,1,0]
Music Mode	The *music* information is not empty, and the *location* is bars, concert halls, coffee shops, residential areas, or other entertainment and life places.	[1,0,1,1,1]
Emotion Mode	Only *facial expression* information is provided.	[0,0,1,0,0]
Weather Mode	Only *weather* information is provided.	[0,0,0,1,0]



Input (Contextual information collected by APIs)
User Command: [Please recommend painting topics based on my current content]
Location: [A bridge above the lake]
Screen Content: [Chat with a friend]
Facial Expression: [Happiness]
Weather: [Sunny]
Music: [Counting Stars (OneRepublic)]

Standard Vector [1,0,1,1,1] (The screen content is considered irrelevant.)

Vice Vector [0,0,1,1,1] (Based on the Standard Vector, it excludes the location information.)

Extracted Context for Topic 1,2
User Command: [Please recommend painting topics based on my current content]
Location: [A bridge above the lake]
Facial Expression: [Happiness]
Weather: [Sunny]
Music: [Counting Stars (OneRepublic)]

Extracted Context for Topic 3
User Command: [Please recommend painting topics based on my current content]
Facial Expression: [Happiness]
Weather: [Sunny]
Music: [Counting Stars (OneRepublic)]

Figure 5: An actual example of Context Selector.

system. The description of user intention modes reflects the relationship between relevant contextual data and the user's intent for creating the painting, highlighting the features that are important within the current mode. Context Selector's **planning** component executes two steps. First, it checks if the user's mode matches the predefined user intention modes. If not, this agent autonomously defines a new mode. Then, it generates a corresponding "standard vector". To ensure robust and diverse topic recommendations, we introduce a "vice vector", which slightly changes the "standard vector", randomly adding or removing information from a specific modality. Table 3 provides the system prompt for Context Selector, and Figure 5 shows an actual example of Context Selector.

4.2.2 "Framing" Phase - Topic Agent. Once Context Selector generates the "standard vector" and "vice vector", it provides the relevant context to **Topic Agent** (Figure 5). Then, Topic Agent presents the user with three theme recommendations with LLM. The first two are derived from the "standard vector", while the third is based on the vice vector. After the user chooses a specific topic or creates a custom topic, Topic Agent will provide a detailed ideation paragraph around the topic.

4.2.3 "Focusing" Phase - Tool Manager. Our formative study discovered a conflict between users' need for high image quality and preference for a low interaction burden. In response, we crafted a conversational interface to address this conflict, which allows preset response selections, intuitive image generation and modification through typed or voice commands. To support the conversational image generation and editing, we present **Tool Manager**, which divides into the following components:

- **Planning**: First, Tool Manager analyzes **memory** (e.g., conversation history) to decide the most appropriate model. Then, the agent crafts prompts for high-quality image generation and editing.

- **Memory**: Tool Manager keeps track of the conversation history, which helps Tool Manager remember past interactions, preferences, and other relevant data in the current conversation. This accumulated knowledge aids in crafting effective prompts. Furthermore, We provide a tool database in the system prompt as the internal memory (Appendix C, Table 6).
- **Tool use**: It opts for image generation with image generative models (e.g., *Stable Diffusion*) and chooses the most appropriate model for image modification (e.g., *ControlNet* models, which can add additional conditions to control the outcome of diffusion models). For instance, Alex mentioned "Bridge," expressing a desire to render a bridge in the original Van Gogh style. Tool Manager would choose the most appropriate *Shuffle* model within *ControlNet*, which utilizes a random flow for reordering the original image. This model then leverages *Stable Diffusion* to transfer the style to a fresh image.

4.2.4 *"Focusing" Phase - Artist Agent.*

Prior work found that LLMs can effectively assist users by inspiring ideas and promptly suggesting specific image content, such as primary elements, composition, style, and color schemes [33, 34, 70]. Deploying LLM as a creative artist persona can provide users with a better co-creation experience [11]. Thus, we present **Artist Agent**, which uses LLM to generate detailed creative suggestions during discussions, and an image-to-text model (e.g., *VisualGLM-6B*) to describe the generated image. Furthermore, users can proactively seek advice from Artist agent (e.g., by clicking the "Advice" button on the shortcut menu). This reduces users' interaction burden and potentially provides new inspirations.

4.2.5 *Personalization Agent.*

We introduce **Personalization Agent** to enhance personalized recommendations. After the user and ContextCam co-create an image that the user is satisfied with, or at the end of a dialogue, Personalization Agent transforms the dialogue history into a concise summary using LLM. This summary includes the key preferences of the user, such as preferred subjects, styles, and themes. In future dialogues, Topic Agent and Artist Agent will refer to this personalized information for recommendations.

4.3 Implementation

We implemented ContextCam as a mobile painting assistant for Android (Figure 2). The conversational interface of this assistant appears on the mobile screen as a resizable floating window. Users can upload photos and input text by typing or through voice recognition to start a conversation. The mobile painting assistant collects five types of contextual data in real-time with the help of different web APIs: location, screen content (text displayed on screen and active application), facial expression (facial expression recognition results from the front-facing camera), weather, and music (music title and artist)[2]. A remote server analyzes these data and responds to the mobile assistant.

We used the GPT-4 API [43] for understanding semantics and generating creative ideas, and employed VisualGLM-6B [60] for image narratives. ContextCam employed zero-shot techniques [45] to craft prompts for ControlNet [36] and Stable Diffusion V1.5 [52].

5 USER STUDY

To evaluate the performance of our system in practical settings and understand user behavior patterns within it, we conducted a real-world user study. After each conversation, we gathered satisfaction ratings. Following the experiment, we developed a questionnaire to gauge user feedback on the system's inspiration, engagement, enjoyment, and usability. Subsequently, we conducted interviews with the participants.

5.1 Participants

The researchers distributed recruitment messages in student group chats and encouraged students to repost these messages on social media platforms, and recruited 16 participants (8 male participants, 8 female participants). Participants (P1-P16) aged between 18 and 46 (avg=24.6, SD=9.60)[3]. All participants received compensation for their time. We obtained ethical approval and participant consent before the study.

All participants had heard of AI image generation, but few participants (5/16) had hands-on experience with AI image generation tools. The main reasons for not using these tools included a lack of need for AI-generated images (6/11), the complexity of the image generators (2/11), and low output quality (3/11).

5.2 Procedure

The study included a 30-minute tutorial session, a three-day real-world usage session followed by a questionnaire, and a 30-minute interview session. Researchers and participants communicated remotely through a social media application and video conferencing software during this process.

5.2.1 *Tutorial Session.*

Prior to the study, we arranged a 30-minute tutorial session for each participant. In this session, we clarified the basic concepts of context awareness, our user study goals, specific tasks, and requirements. Then, we introduced ContextCam in detail and demonstrated its features.

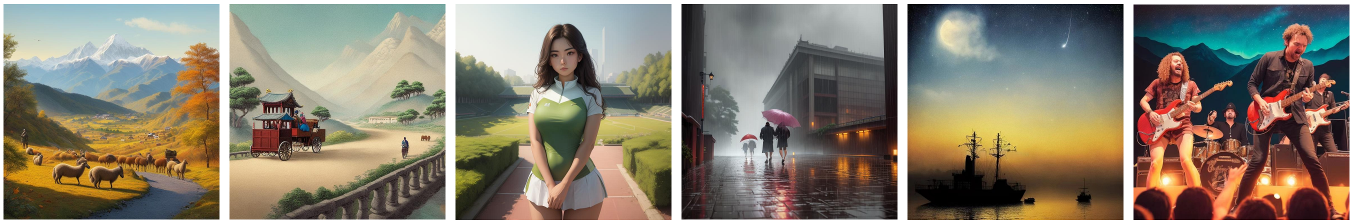
5.2.2 *Real-World Context-Aware Image Creation with ContextCam.*

In this session, 16 participants used ContextCam in real-world scenarios for three days. The task required about five attempts per day, totaling 15 attempts. The researchers sent daily messages to encourage participants to explore their creative potential and experiment with ContextCam in a variety of scenarios to enhance the diversity of the experiment. Each participant engaged in 15 conversations, and participants were involved in 136 unique scenarios in total. To evaluate the abilities of ContextCam and analyze user interaction, we collected interaction data such as users' dialogue records, generated images, and contextual information. We conducted thematic analysis [5] on the interaction data to extract key themes and gain deeper insights into the users' experiences with ContextCam. After each conversation, we captured immediate feedback on users' satisfaction ratings with the generated images. Users offered reasons when dissatisfied.

---

[2] A complete list of APIs used can be found in Appendix A.

[3] The details of participants are in the supplementary material.



(a) "Herding on the highlands in autumn." (P5)

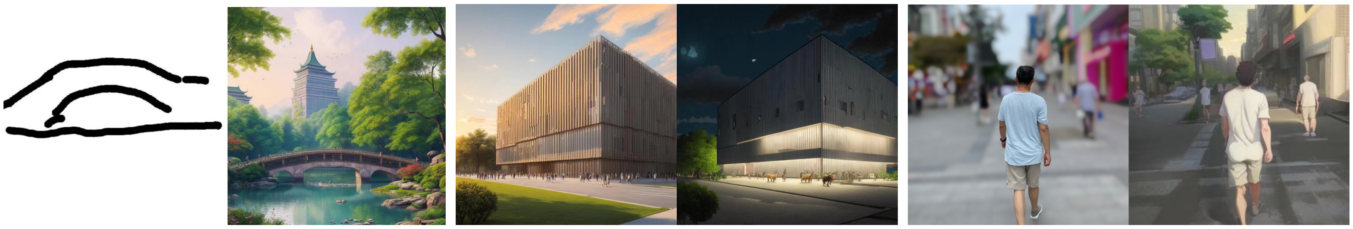
(b) "Princess Wencheng's journey to Tibet." (P5)

(c) "A college girl in a sports skirt." (P9)

(d) "A classroom building on a rainy day." (P9)

(e) "Sunset over the sea." (P10)

(f) "A rock band." (P8)



(g) Sketch Input and Result ("Bridge") (P8)

(h) "Night. Turn people into animals." (P2)

(i) "Cartoon style." (P1)

Figure 6: Some images from the experiment that users found satisfactory. The themes of (a)(d)(e) are entirely recommended based on the context. (b)(c)(f) are the results of recommendations that combine user commands with the current context. (g) is an example of using contextual data to transform a simple sketch into a beautiful landscape painting. (h) and (i) show images before and after editing only based on user commands.

5.2.3 Survey and Interview. After all conversations, a questionnaire[4] was presented to understand the user's overall experience with ContextCam. Participants rated their level of enjoyment, degree of facilitation of inspiration by ContextCam, ContextCam's usability, and sense of collaboration with ContextCam on a 7-point Likert scale (1=strongly disagree, 7=strongly agree)[5]. We utilized content analysis [30] to categorize and interpret the textual data from user feedback both after each conversation and all conversations, which helped us understand the system's limitations and users' overall experience.

Then, we conducted interviews for each participant around three themes: 1) context-driven behavior patterns within ContextCam, 2) context awareness and human-AI co-creation, and 3) contextual information as an inspiration and its impact. The interview for each participant lasted approximately 30 minutes. With the consent of participants, we collected 8 hours of audio recordings, which were subsequently transcribed for analysis. One researcher conducted open coding based on grounded theory methodology [20] to uncover insights related to these three themes.

6 RESULTS AND DISCUSSION

First, we considered three aspects for evaluating ContextCam's performance: user satisfaction with images (section 6.1), ContextCam's ability to infer user intentions (section 6.2), and interaction burden (section 6.3). Then, we delved into users' interaction behaviors with ContextCam and their feedback to uncover insights about the behavior patterns demonstrated by the users (section 6.4), how context awareness affected human-AI co-creation (section 6.5), and how users used contextual information as inspiration (section 6.6). Figure 6 showcases select works from our user study, while Figure 7 displays the post-experiment questionnaire ratings from participants.

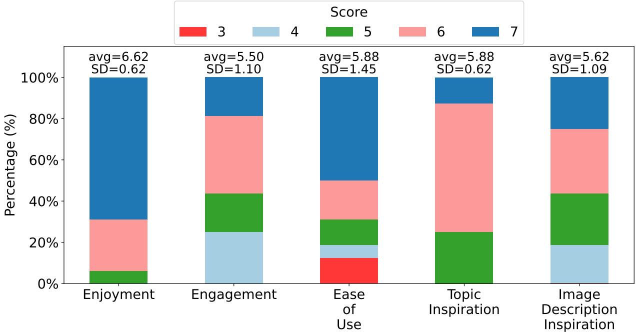


Figure 7: The stacked bar chart presents the user evaluation results for ContextCam, using the 7-point Likert scale, with no users assigning low scores of 1 or 2.

6.1 User Satisfaction Rating for Images

After each conversation, we collected self-reported user satisfaction with the images on the 7-point Likert scale. Figure 8 shows the user satisfaction ratings. The average user rating was 5.80 (SD=1.17),

---

[4] All survey and interview questions can be found in the supplementary material.

[5] Throughout this article, when we refer to the 7-point Likert scale, it is understood to range from 1 (strongly disagree) to 7 (strongly agree).

indicating a relatively high level of satisfaction. In a total of 240 conversations, 154 conversations (64%) had a satisfaction score above 5 (score=6or7), and 12 conversations (5%) had a satisfaction score below 4 (score=2or3), which we perceived as unsatisfied conversations. Five unsatisfied conversations (2%) were due to AI image generation defects in characters' hands or feet. Aside from this, we categorized other common reasons for unsatisfied conversations:

1. **Cultural Mismatch in Aesthetics (N=3, 1.25%):** Cultural differences shape user aesthetic expectations. For instance, a Western theater style clashed with Eastern aesthetics (P9, Figure 9(a)).
2. **Divergence in Scene Perception (N=2, 0.83%):** There is a gap between how the diffusion model depicts certain scenes and user perceptions. For example, the school playground depicted by the model differed from P5's familiar scene (Figure 9(b)).
3. **Model's Limited Knowledge (N=2, 0.83%):** The diffusion model sometimes lacks knowledge of specific locations, dishes, or characters (e.g., P4, Figure 9(c)).

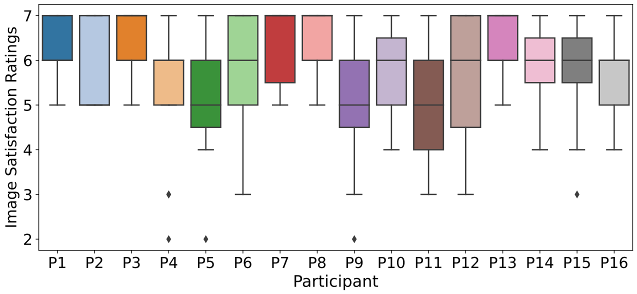
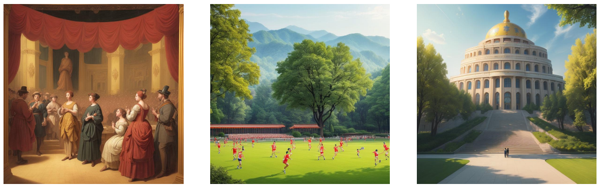


Figure 8: The boxplot illustrates users' satisfaction ratings for the images, evaluated on the 7-point Likert scale.[6]



(a) "I thought of a Chinese theater but drew a Western one." (P9)

(b) "This playground differs from the one I'm familiar with." (P5)

(c) "It doesn't know my university." (P4)

Figure 9: Failure cases from the user study.

6.2 User Intent Inference

In each topic recommendation message, ContextCam provides three topics. Topics 1 and 2 use the same user contextual data generated by Context Selector, while Topic 3 randomly incorporates or omits a piece of contextual information to enhance robustness. When users are dissatisfied with all suggestions, they can directly type in a desired topic as a custom topic.

To evaluate the ability of ContextCam to infer user intentions, we examined the percentage of conversations where users selected recommended topics and conversations where they proposed custom topics. Figure 10 displays the distribution of users' final theme selections. For all conversations, users picked system-generated topics (topics 1, 2, or 3) 92.9% of the time.

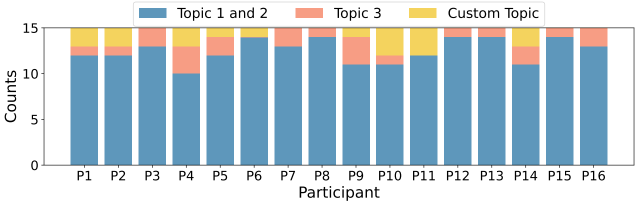


Figure 10: Stacked bar chart of the counts of each user's choice of topic.

6.3 Interaction Burden

We define a "conversation" as a user's complete discussion with ContextCam on a specific topic, encompassing multiple "interactions" and one or more images. An "interaction" refers to the user sending one message (feedback on image themes, proposals, image generation commands, or editing instructions) and receiving one response from ContextCam. The term "number of input words" refers to the count of Chinese characters.

To evaluate the user's interaction burden with ContextCam, we calculated the number of input words (Figure 11(a)), clicks (Figure 11(c)), and interaction rounds (Figure 11(c)) for each conversation. Then, we calculated the ratio of input words to interaction rounds for each conversation (Figure 11(b)). On average, this ratio stood at 1.1 for a single interaction, indicating a low interaction burden. In the post-experiment user survey, the average usability rating was 5.88 (SD=1.45) on the 7-point Likert scale (Figure 7).

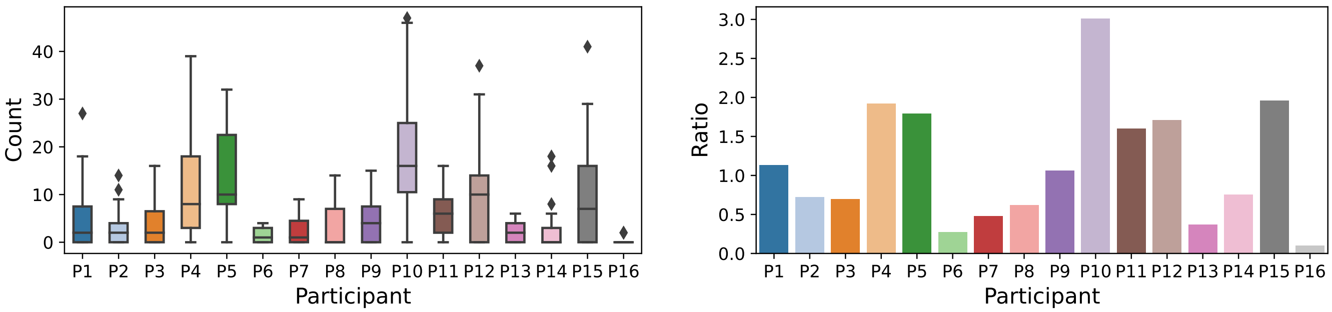
6.4 Behavior Patterns within ContextCam

Participants varied in their frequency and preference for using different types of contextual information (Figure 12). Factors such as traveling events, lifestyle, and personality influenced the results. For instance, P13 was on a road trip and used location data in 13 of 15 conversations. In contrast, P12 and P16, who preferred staying at home, used location data only 3 and 4 times, respectively. Two participants (P6, P11) stated they preferred not

---

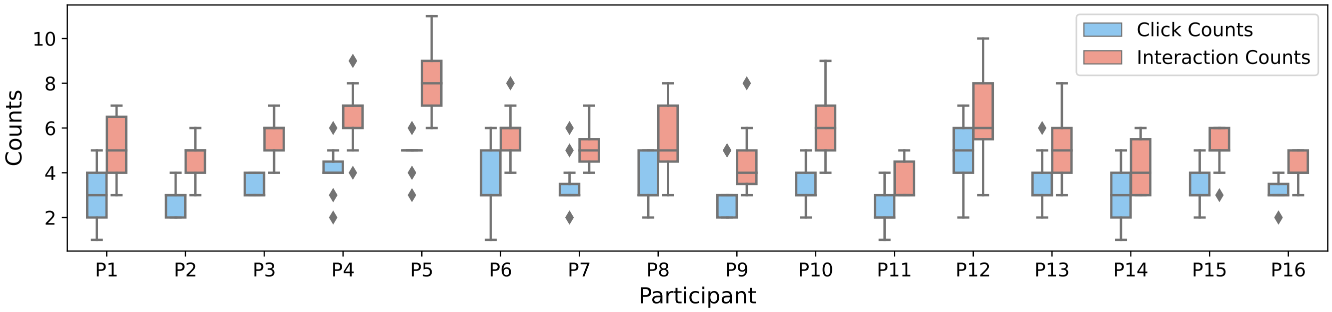
[6] The central line within the box denotes the median, while the upper and lower edges correspond to the third and first quartiles, respectively. The whiskers capture the range of the data, excluding outliers. Diamonds in the graph signify outliers that deviate from the typical interquartile range.

[7] In a boxplot (Figure 12(a)(c)), the central line within the box denotes the median, while the upper and lower edges correspond to the third and first quartiles, respectively. The whiskers capture the range of the data, excluding outliers. Diamonds in the graph signify outliers that deviate from the typical interquartile range.



(a) Total word count per conversation.

(b) Average word count per interaction.



(c) Click counts and interaction counts per conversation.

Figure 11: Interaction burden[7].

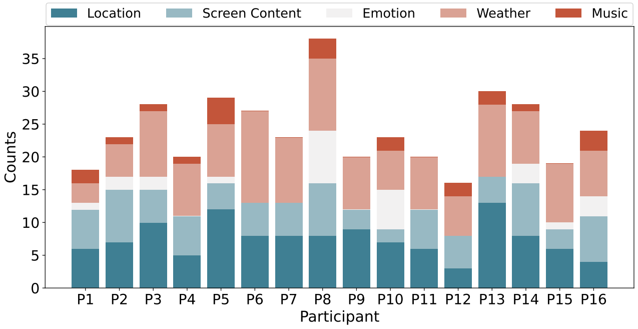


Figure 12: Counts of types of contextual data used in conversations.

to use music information because they typically used ContextCam in quiet settings and did not play music out loud.

Participants enjoyed using contextual information to create images. Out of 240 user conversations, in only four conversations (1.7%, P1, P7, P14, P16), users decided not to include any contextual data, citing reasons such as "the current context is completely irrelevant to my creative intention" and "I already had a specific theme in mind beforehand." Nine participants each used the shortcut command "Please recommend painting topics based on my current context," 2 or 3 times to fully adopt their immediate context as the image theme. P15, who had extensive experience with AI image generators, commented: "Integrating context makes AI image generation more interesting and personalized. It opens up my creative thought, much like sketching from life."

Contextual information affected users' image theme preferences. P9 frequently created images based on various scenes from his university life, with themes centered around campus experiences such as "a college girl in a sports skirt" (Figure 6(c)), "convert a lab picture into a sci-fi style," and "a classroom building on a rainy day." (Figure 6(d)) While on a road trip, P5 often created images with views along the route (Figure 6(a)), local culture and history (Figure 6(b)).

Users frequently weaved in their personal experiences and preferences. For example, when users created art under specific weather conditions, they may unconsciously associate their moods and scenes with the weather. Sunny days may suggest happiness, excitement, and vitality, while cloudy days may suggest sadness, meditation, and relaxation. P1 said, "Most of the time, it was raining when I conducted my experiments, so the themes of my creations often had a sad tone." Locations can also inspire related cultural themes. Consider P5, who selected certain iconic venues during a journey, such as historic cities, scenic landscapes, or region-specific dwellings, to represent specific cultural themes. These locations are often rich in cultural connotations and unique aesthetic values, which can inspire users and add more cultural flavor to their works.

As learners gained familiarity with ContextCam, they were more likely to actively customize and leverage context awareness in image creation. Initially, users (P1, P4, P8-P10, P13-P16) primarily explored the context by leveraging shortcuts 2 or 3 times. As their familiarity with the contextual data grew, their methods of using this data for creation diversified. All 16 participants proactively input simple themes, sketches, or photographs over 8 times, prompting AI to create based on the context. Through this, not only did users acquire a deeper grasp of context-aware co-creation, but they also developed a unique understanding of the contextual data's intrinsic qualities and potential value. Building on this understanding, users can adjust various elements flexibly, leading to efficient, precise, and innovative outputs. As users delved deeper into context-aware co-creation, some selectively turned off certain data inputs based on their preferences, such as location and screen content information.

In the long run, these insights hold significant implications for advancing and innovating in context-aware co-creation. As Figure 7 shows, users provided positive feedback on the enjoyment, engagement, and inspiration of ContextCam.

6.5 Context Awareness and Human-AI Co-Creation

Context awareness made the collaborative process more enjoyable and efficient, resulting in distinctively personalized outputs. Here, creation is no longer an isolated behavior; it is co-shaped by external environments, emotions, and interactions between users and AI. Context awareness offers a backdrop for human-AI image co-creation, guiding the themes and directions of the process. P10 mentioned, "When I was at the seaside, ContextCam recommended the theme 'Sunset over the sea.' This really sparked my imagination, and I started incorporating additional elements like 'sailboat' and 'meteor,' enriching the original idea." (Figure 6(e))

Context awareness fostered emotional resonance and encouraged users' engagement in human-AI co-creation. Immersing users within a relevant context increased their emotional and cognitive connection to the content. This was evident from our survey: in Figure 7, users reported an average engagement rating 5.50 (SD=1.10). Moreover, their inspiration derived from image descriptions was impressively rated at 5.62 (SD=1.09), suggesting that the context greatly improved the overall user experience. P8 mentioned, "Using ContextCam made me feel more connected to the surroundings. I opened ContextCam next to an old-fashioned tower. The sky was clear, with a few clouds drifting by. I simply sketched the shape of a bridge, but ContextCam perfectly captured that moment." (Figure 6(g))

6.6 Contextual Information as an Inspiration

Users perceived an enhancement in their creative processes due to context awareness, and the themes recommended by ContextCam amazed them. ContextCam employed contextual data to create images, vividly showcasing aspects of users' local natural environment (P5, Figure 6(a)), cultural landmarks (P8, Figure 6(g)), and historical context (P5, Figure 6(b)) and so on. P1 reported, "As I get the hang of ContextCam, I feel like my creative boundaries are expanding, and I'm inspired to make deeper and more meaningful creations. I'm no longer limited to my imagination, and I've never been more excited to create." These themes enriched their inspiration and deepened their understanding of using contextual information. For instance, when P8 aimed to depict "mountains under the night sky," ContextCam suggested "a rock band performing with mountains under the night sky as the background" (Figure 6(f)), cleverly linking the theme with the rock music the user was currently playing. Consequently, P8 frequently incorporated music into subsequent creations.

7 LIMITATIONS AND FUTURE WORK

This paper presents the preliminary application of context awareness in human-AI image co-creation. We highlight the transformative potential of AIGC in documenting life and self-expression, particularly within the paradigm of context-aware human-AI co-creation. This section introduces the key directions ContextCam aims to explore in its future development.

7.1 Deepening Context Awareness

ContextCam extracts five types of contextual information using sensors and efficiently filters out irrelevant data via the Few-Shot-CoT method, enhancing the understanding of user intent. Our study represents a preliminary exploration. As multi-modal models continue to develop, so will the capabilities of ContextCam. ContextCam may incorporate more diverse data types like motion states and scents to comprehensively reflect the user's real-world context. One can envision a day when the aroma of freshly baked cookies sparks creative inspiration.

With the advancements in sensing technology, we will incorporate real-time detection of user physiological data, including heart rate, body temperature, activity levels, sleep quality, and step count [13]. In parallel, environmental sensing, capturing metrics such as external temperature, lighting conditions, humidity, and air quality, are also research priorities. By synthesizing this data, we can comprehensively understand a user's health, activity status, and environment, producing more personalized and engaging images. Other advanced technologies, such as Brain-Computer Interfaces (BCI) [64], present new avenues to explore user intentions and the subconscious and can enhance the potential for image creation.

Understanding the user's sociocultural background is also crucial to thoroughly achieving context awareness. We will explore ways to integrate this background information as a vital component of contextual presentation.

7.2 Optimizing Co-Creation and Inspiring Creativity

Because of user diversity, we plan to introduce advanced personalization tools, such as personalized knowledge graphs, ensuring a unique experience for each user. We will also proactively integrate various external information resources to prevent data silos. Moreover, recognizing that every user has unique behavioral patterns and preferences, we aim to offer more adaptable guidance strategies. This helps them explore new ideations, enhancing their overall experience. In the future, ContextCam will develop diverse motivational mechanisms and features to inspire user creativity and boost engagement.

7.3 Innovating Interaction Modes

In the future, ContextCam can be integrated into wearable devices such as smartwatches and wristbands. These devices' rich sensor data and unique interactive features pave the way for expanding our system's capabilities. Additionally, we are confident of ContextCam's significant potential in Augmented Reality (AR) and Virtual Reality (VR) domains. Integrating it into AR and VR platforms can offer users a more immersive and intuitive experience. While we have made promising strides, the potential avenues for expansion and enhancement are vast. As technology evolves, the combination of human intelligence and AI capabilities will keep redefining the boundaries of creativity.

8 CONCLUSION

This paper proposed a context-aware human-AI co-creation system, ContextCam. To assess its effectiveness, we conducted a study involving 16 participants and 136 scenarios. Results indicated that users picked system-generated topics in 92.9% of the cases. Users also expressed high satisfaction with the final images (avg=5.80, SD=1.17) and enjoyed the creative process (avg=6.62, SD=0.62). Detailed interviews and analyses of user interactions with ContextCam revealed some findings about how contextual data is utilized and perceived during the creative process. These findings showed how contextual information shapes image themes, influences user behaviors, and enhances the collaborative experience between humans and AI through ContextCam. Overall, ContextCam's approach to context-aware co-creation is perceived as enjoyable, integrating contextual information to boost creativity, engagement, and inspiration. Although ContextCam has received positive feedback from users, there are challenges to address in real-world applications, such as the diversity in contextual extraction and system response times. Future work will focus on refining ContextCam and exploring new application scenarios to advance the field of context-aware human-AI image co-creation.

REFERENCES

[1] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2021. Context-Aware Target Apps Selection and Recommendation for Enhancing Personal Mobile Assistants. *ACM Trans. Inf. Syst.* 39, 3, Article 29 (may 2021), 30 pages. https://doi.org/10.1145/3447678

[2] AUTOMATIC1111. 2022. Negative prompt. https://github.com/AUTOMATIC1111/stable-diffusion-webui/wiki/Negative-prompt.

[3] AUTOMATIC1111. 2023. stable-diffusion-webui. https://github.com/AUTOMATIC1111/stable-diffusion-webui GitHub repository.

[4] Seungho Baek, Hyerin Im, Jiseung Ryu, Juhyeong Park, and Takyeon Lee. 2023. PromptCrafter: Crafting Text-to-Image Prompt through Mixed-Initiative Dialogue with LLM. arXiv:2307.08985 [cs.HC]

[5] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.

[6] Gabriel Caniglia. 2020. Cast: A Context-Aware Collaborative Storytelling Platform. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3334480.3382966

[7] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun. 2023. A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. arXiv:2303.04226 [cs.AI]

[8] John Carbone and James Crowder. 2014. Collaborative Shared Awareness: Human-AI Collaboration.

[9] Annie Chen. 2005. Context-Aware Collaborative Filtering System: Predicting the User's Preferences in Ubiquitous Computing. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (Portland, OR, USA) *(CHI EA '05)*. Association for Computing Machinery, New York, NY, USA, 1110–1111. https://doi.org/10.1145/1056808.1056836

[10] 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 574, 14 pages. https://doi.org/10.1145/3411764.3445089

[11] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Visual Sketching of Story Generation with Pretrained Language Models. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 172, 4 pages. https://doi.org/10.1145/3491101.3519873

[12] Fabio Clarizia, Francesco Colace, Massimo De Santo, Marco Lombardi, Francesco Pascale, and Domenico Santaniello. 2019. A Context-Aware Chatbot for Tourist Destinations. In *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. 348–354. https://doi.org/10.1109/SITIS.2019.00063

[13] Sunny Consolvo, David W. McDonald, Tammy Toscos, Mike Y. Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, Ian Smith, and James A. Landay. 2008. Activity Sensing in the Wild: A Field Trial of Ubifit Garden. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 1797–1806. https://doi.org/10.1145/1357054.1357335

[14] Anind K. Dey. 2001. Understanding and Using Context. 5, 1 (jan 2001), 4–7. https://doi.org/10.1007/s007790170019

[15] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A Survey on In-context Learning. arXiv:2301.00234 [cs.CL]

[16] Sergio Ferrando and Eva Onaindia. 2013. Context-Aware Multi-Agent Planning in intelligent environments. *Information Sciences* 227 (04 2013), 22–42. https://doi.org/10.1016/j.ins.2012.11.021

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

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

[19] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind One Embedding Space to Bind Them All. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15180–15190. https://doi.org/10.1109/CVPR52729.2023.01457

[20] Barney Glaser and Anselm Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Aldine Transaction.

[21] 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* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 226, 20 pages. https://doi.org/10.1145/3544548.3580999

[22] Xu Haoran, Chen Shuyao, and Ying Zhang. 2023. Magical Brush: A Symbol-Based Modern Chinese Painting System for Novices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 131, 14 pages. https://doi.org/10.1145/3544548.3581429

[23] Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, and Chenglin Wu. 2023. MetaGPT: Meta Programming for Multi-Agent Collaborative Framework. arXiv:2308.00352 [cs.AI]

[24] Mina Huh, Yi-Hao Peng, and Amy Pavel. 2023. GenAssist: Making Image Generation Accessible. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 38, 17 pages. https://doi.org/10.1145/3586183.3606735

[25] Francisco Ibarrola, Tomas Lawton, and Kazjon Grace. 2023. A Collaborative, Interactive and Context-Aware Drawing Agent for Co-Creative Design. *IEEE Transactions on Visualization and Computer Graphics* (2023), 1–13. https://doi.org/10.1109/TVCG.2023.3293853

[26] Jinglu Jiang, Alexander J. Karran, Constantinos K. Coursaris, Pierre-Majorique Léger, and Joerg Beringer. 2021. A Situation Awareness Perspective on Human-Agent Collaboration: Tensions and Opportunities. In *HCI International 2021 - Late Breaking Papers: Multimodality, EXtended Reality, and Artificial Intelligence: 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings*. Springer-Verlag, Berlin, Heidelberg, 437–444. https://doi.org/10.1007/978-3-030-90963-5_33

[27] Pegah Karimi, Jeba Rezwana, Safat Siddiqui, Mary Lou Maher, and Nasrin Dehbozorgi. 2020. Creative sketching partner: an analysis of human-AI co-creativity. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 221–230.

[28] Shahedul Huq Khandkar. 2009. Open Coding. *University of Calgary* 23 (2009), 2009.

[29] Jeongyeon Kim, Yubin Choi, Meng Xia, and Juho Kim. 2022. Mobile-Friendly Content Design for MOOCs: Challenges, Requirements, and Design Opportunities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 92, 16 pages. https://doi.org/10.1145/3491102.3502054

[30] Klaus H. Krippendorff. 2003. *Content Analysis: An Introduction to Its Methodology (2nd ed.)*. Sage Publications, Inc.

[31] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Scale Language Model Society. arXiv:2303.17760 [cs.AI]

[32] Vivian Liu, Tao Long, Nathan Raw, and Lydia Chilton. 2023. Generative Disco: Text-to-Video Generation for Music Visualization. arXiv:2304.08551 [cs.HC]

[33] Vivian Liu, Han Qiao, and Lydia Chilton. 2022. Opal: Multimodal Image Generation for News Illustration. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 73, 17 pages. https://doi.org/10.1145/3526113.3545621

[34] 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* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 1955–1977. https://doi.org/10.1145/3563657.3596098

[35] Xiaoyi Liu, Yingtian Shi, Chun Yu, Cheng Gao, Tianao Yang, Chen Liang, and Yuanchun Shi. 2023. Understanding In-Situ Programming for Smart Home Automation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 2, Article 66 (jun 2023), 31 pages. https://doi.org/10.1145/3596254

[36] lllyasviel. 2023. ControlNet. https://huggingface.co/lllyasviel/ControlNet.

[37] Andrés Lucero. 2015. Using Affinity Diagrams to Evaluate Interactive Prototypes. In *IFIP Conference on Human-Computer Interaction*. Springer, 231–248.

[38] Midjourney. 2023. Midjourney. https://www.midjourney.com Accessed: May 23, 2023.

[39] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. 2023. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. arXiv:2302.08453 [cs.CV]

[40] Elizabeth L. Murnane, Yekaterina S. Glazko, Jean Costa, Raymond Yao, Grace Zhao, Paula M. L. Moya, and James A. Landay. 2023. Narrative-Based Visual Feedback to Encourage Sustained Physical Activity: A Field Trial of the WhoIsZuki Mobile Health Platform. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 1, Article 23 (mar 2023), 36 pages. https://doi.org/10.1145/3580786

[41] Yugo Nakamura, Rei Nakaoka, Yuki Matsuda, and Keiichi Yasumoto. 2023. Eat2pic: An Eating-Painting Interactive System to Nudge Users into Making Healthier Diet Choices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 1, Article 24 (mar 2023), 23 pages. https://doi.org/10.1145/3580784

[42] OpenAI. 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt.

[43] OpenAI. 2023. GPT-4 API. https://openai.com/gpt-4.

[44] Hiroyuki Osone, Jun-Li Lu, and Yoichi Ochiai. 2021. BunCho: AI Supported Story Co-Creation via Unsupervised Multitask Learning to Increase Writers' Creativity in Japanese. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 19, 10 pages. https://doi.org/10.1145/3411763.3450391

[45] Mark Palatucci, Dean Pomerleau, Geoffrey Hinton, and Tom M. Mitchell. 2009. Zero-Shot Learning with Semantic Output Codes. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems* (Vancouver, British Columbia, Canada) *(NIPS'09)*. Curran Associates Inc., Red Hook, NY, USA, 1410–1418.

[46] Balaji Parasumanna Gokulan and D. Srinivasan. 2010. *An Introduction to Multi-Agent Systems*. Vol. 310. 1–27. https://doi.org/10.1007/978-3-642-14435-6_1

[47] Louisa Penfold. 2017. Art as experience. *Educational Review* 69, 4 (2017), 523–523. https://doi.org/10.1080/00131911.2016.1264206 arXiv:https://doi.org/10.1080/00131911.2016.1264206

[48] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952 [cs.CV]

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[50] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125 [cs.CV]

[51] 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 (CVPR)*. 10684–10695.

[52] runwayml. 2023. Stable Diffusion V1.5. https://huggingface.co/runwayml/stable-diffusion-v1-5.

[53] Murray Shanahan. 2023. Talking About Large Language Models. arXiv:2212.03551 [cs.CL]

[54] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. arXiv:2303.17580 [cs.CL]

[55] Jaidev Shriram and Sanjayan Pradeep Kumar Sreekala. 2023. ZINify: Transforming Research Papers into Engaging Zines with Large Language Models. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, Article 117, 3 pages. https://doi.org/10.1145/3586182.3625118

[56] George A. Sielis, Aimilia Tzanavari, and George A. Papadopoulos. 2009. Enhancing the Creativity Process by Adding Context Awareness in Creativity Support Tools. In *Proceedings of the 5th International Conference on Universal Access in Human-Computer Interaction. Part III: Applications and Services* (San Diego, CA) *(UAHCI '09)*. Springer-Verlag, Berlin, Heidelberg, 424–433. https://doi.org/10.1007/978-3-642-02713-0_45

[57] Sarah Sterman, Evey Huang, Vivian Liu, and Eric Paulos. 2020. Interacting with Literary Style through Computational Tools. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376730

[58] Xiang Zhi Tan, Samantha Reig, Elizabeth J. Carter, and Aaron Steinfeld. 2019. From One to Another: How Robot-Robot Interaction Affects Users' Perceptions Following a Transition Between Robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 114–122. https://doi.org/10.1109/HRI.2019.8673304

[59] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Agüera y Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. *CoRR* abs/2201.08239 (2022). arXiv:2201.08239 https://arxiv.org/abs/2201.08239

[60] THUDM. 2023. VisualGLM-6B. https://github.com/THUDM/VisualGLM-6B.

[61] Ziqi Wang, Bin Guo, Hao Wang, Helei Cui, Yang He, and Zhiwen Yu. 2020. MateBot: The Design of a Human-Like, Context-Sensitive Virtual Bot for Harmonious Human-Computer Interaction. In *Green, Pervasive, and Cloud Computing: 15th International Conference, GPC 2020, Xi'an, China, November 13–15, 2020, Proceedings* (Xi'an, China). Springer-Verlag, Berlin, Heidelberg, 273–287. https://doi.org/10.1007/978-3-030-64243-3_21

[62] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 24824–24837. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

[63] Lilian Weng. 2023. LLM-powered Autonomous Agents. *lilianweng.github.io* (Jun 2023). https://lilianweng.github.io/posts/2023-06-23-agent/

[64] Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. 2002. Brain–computer interfaces for communication and control. *Clinical Neurophysiology* 113, 6 (2002), 767–791. https://doi.org/10.1016/S1388-2457(02)00057-3

[65] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models. arXiv:2303.04671 [cs.CV]

[66] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework. arXiv:2308.08155 [cs.AI]

[67] 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* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 385, 22 pages. https://doi.org/10.1145/3491102.3517582

[68] Mengru Xue, Rong-Hao Liang, Bin Yu, Mathias Funk, Jun Hu, and Loe Feijs. 2019. AffectiveWall: Designing Collective Stress-Related Physiological Data Visualization for Reflection. *IEEE Access* 7 (2019), 131289–131303. https://doi.org/10.1109/ACCESS.2019.2940866

[69] Daijin Yang, Yanpeng Zhou, Zhiyuan Zhang, Toby Jia-Jun Li, and Ray LC. 2022. AI as an Active Writer: Interaction strategies with generated text in human-AI collaborative fiction writing. In *Joint Proceedings of the ACM IUI Workshops*, Vol. 10. CEUR-WS Team.

[70] Chaoning Zhang, Chenshuang Zhang, Sheng Zheng, Yu Qiao, Chenghao Li, Mengchun Zhang, Sumit Kumar Dam, Chu Myaet Thwal, Ye Lin Tun, Le Luang Huy, Donguk kim, Sung-Ho Bae, Lik-Hang Lee, Yang Yang, Heng Tao Shen, In So Kweon, and Choong Seon Hong. 2023. A Complete Survey on Generative AI (AIGC): Is ChatGPT from GPT-4 to GPT-5 All You Need? arXiv:2303.11717 [cs.AI]

[71] 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 (ICCV)*. 3836–3847.

[72] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic Chain of Thought Prompting in Large Language Models. arXiv:2210.03493 [cs.CL]

A IMPLEMENTATION DETAILS

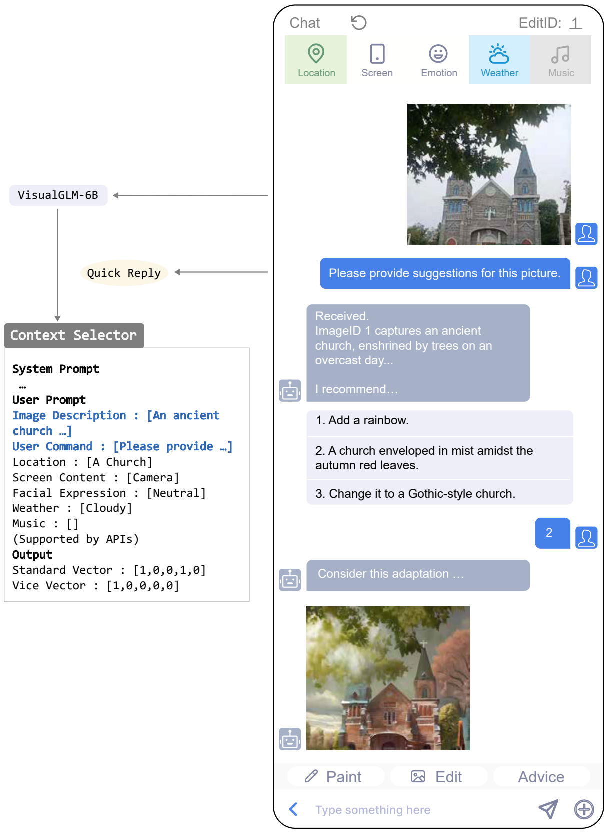
Table 4 outlines the types of context detectors incorporated into the systems, detailing their purpose and the types of data they provide.

Table 4: Context detectors used in the implemented systems.

Context Detector	Description	Value
Location[1]	Using both network and GPS positioning, giving priority to the more accurate location result.	
Screen Content	Capturing the text displayed on the user's current screen and identifying the app being used.	
Facial Expression[2]	Returning facial expression recognition results by capturing the user's frontal face.	e.g., "Happiness," "Anger"
Weather[3]	Retrieving real-time weather data for the user's location.	e.g., "Sunny," "Cloudy"
Music[4]	If music is detected in the environment, returning its song title and the artist.	e.g., "Counting Stars (OneRepublic)"

B ANOTHER USE CASE OF CONTEXTCAM (PHOTO INPUT)

Figure 13 shows users can also take photos and send them to ContextCam.



```
System Prompt
 …
User Prompt
Image Description : [An ancient
church …]
User Command : [Please provide …]
Location : [A Church]
Screen Content : [Camera]
Facial Expression : [Neutral]
Weather : [Cloudy]
Music : []
(Supported by APIs)
Output
Standard Vector : [1,0,0,1,0]
Vice Vector : [1,0,0,0,0]
```

Please provide suggestions for this picture.

Received.
ImageID 1 captures an ancient church, enshrined by trees on an overcast day...

I recommend…

1. Add a rainbow.
2. A church enveloped in mist amidst the autumn red leaves.
3. Change it to a Gothic-style church.

2

Consider this adaptation …

Figure 13: In the scenario, the user with a *neutral expression* photographed a *church* on a *cloudy* sky. Upon receiving the image, ContextCam employed VisualGLM-6B to generate a description of the picture. This description was then conveyed to Context Selector in the format "Image Description: []." Context Selector saw "Location" and "Weather" as pertinent information, subsequently providing three suggestions (standard vector: [1,0,0,1,0], vice vector: [1,0,0,0,0]).

C SYSTEM PROMPTS FOR PERSONALIZATION AGENT AND TOOL MANAGER

Table 5 presents the system prompt and an example for Personalization Agent. The agent accumulates user preferences from each output, and these summaries will be used as part of the system prompts for Topic Agent and Artist Agent in subsequent conversations.

Table 6 displays Tool Manager's prompts. In Stable Diffusion, using a negative prompt tells the model to avoid certain elements or themes in its outputs, which helps improve the image quality [2].

---

[1] Amap location API: https://lbs.amap.com/product/locate/

[2] MEGVII facial expression recognition API: https://www.faceplusplus.com/emotion-recognition/

[3] Dev.QWeather API: https://dev.qweather.com/docs/api/weather/weather-now/

[4] ACRCloud song recognition API: http://console.acrcloud.cn/service/avr

Table 5: System prompt and an actual example for Personalization Agent.

System Prompt	This is the chat history between the user and the painting assistant. Please list keywords to briefly summarize the user's preferred subjects, styles, and themes.
An Actual Example	
Input:	**(Chat History)** User: Please recommend painting topics based on my current context. Assistant: I recommend three painting themes. If you have a better suggestion, please let me know. 1. Emulate Van Gogh's "Starry Night" to depict a tranquil lake on a clear summer night ... 2. Oil Painting: Beneath the delightful clear sky, the band OneRepublic ... 3. A gathering of children joyfully ... User: 1 Assistant: Draw a tranquil lake on a clear summer day in the style of Van Gogh's "Starry Night."... joyful artistic ambiance. User: Generate the image! Assistant: Completed. Image 1 unveils a bridge under the night sky. The heavens are speckled with stars, immersing one in a dreamlike realm... User: Bridge Assistant: Consider this adaptation: Retain the Van Gogh starry night essence in the original image and craft a new piece with a bridge as its focus in the same palette. Capture the sky with...
Output:	
Preferred Style	Van Gogh's "Starry Night" style
Preferred Subject	Tranquil lake, bridge
Preferred Theme	Summer night, clear sky, dreamlike ambiance

Table 6: System prompt for Tool Manager, including tool selection and prompt generation for Stable Diffusion and ControlNet.

Tool Selection	
Choose the most appropriate image modification tool based on the previous discussion and OUTPUT THE NUMBER (1-6):	
Tool	**Description**
1. Shuffle	APPLY the STYLE of the input image to a new image.
2. Softedge_hed	Generate new images without adding or replacing objects/backgrounds from the image. For example, transitioning from day to night or from spring to summer; also involve CHANGING the artistic STYLE, including science fiction, oil painting, watercolor, impressionism, etc.
3. Depth	Replace objects in the image.
4. Openpose	Create a new image with the SAME POSE as the person in the original image.
5. Mlsd	Generate ARCHITECTURAL or INTERIOR DESIGN drawings based on the original image.
6. Canny	Add/Replace/Enrich the background to the picture. Add objects.
Prompt Generation	
Stable Diffusion	
Positive Prompt Generation	ONLY WRITE ENGLISH PROMPT. Give you art discussions between the user and the artist. Use the FINAL RESULT of the discussion. If the user believes the artist's image description is incorrect, you should comply with the user's request. Place the painting theme the user chose at the beginning and write an English prompt for the text-to-image model to draw a picture WITHIN 50 WORDS. Note that if the description is relatively long, you need to extract the central imagery and scenes; if short, emphasize the subject of the painting, employ your imagination, and add some content to enrich the details. DON'T begin with words like "create" or "paint," directly describing the scene.
Negative Prompt Generation	ONLY WRITE ENGLISH PROMPT. You are provided with an art discussion between the user and the artist. Use the FINAL RESULT of the discussion. If the user mentions the people, objects, scenes, or styles they wish to paint, summarize the antonyms of what they want to paint into ENGLISH KEYWORDS, not exceeding six words. If the user does not specify what they do not want to paint, reply with a space. For instance, if the user does not want to paint nighttime, your response should be "night scene"; if the user wants to paint nighttime, your response should be "daytime." DON'T start with words like "create" or "paint."
ControlNet	(We do not generate negative prompts; just use the existing template of negative prompt.)
Positive Prompt Generation	ONLY WRITE ENGLISH PROMPT. You are to receive an art discussion between a user and an artist. Use the FINAL RESULT of the discussion. You need to depict the SCENE of the NEW IMAGE from these perspectives as an ENGLISH PROMPT for the text-to-image model: main characters or objects; background objects; style. After the art discussion, summarize the improvements to the image but retain parts of the original image that were not modified. The prompt should NOT EXCEED 50 WORDS nor include terms like "high contrast."
Prompt Template	
Positive Prompt	((masterpiece, best quality, ultra-detailed, illustration)) + [LLM-generated positive prompt]
Negative Prompt	NSFW, (EasyNegative:0.8), (badhandv4:0.8), (missing fingers, multiple legs), (worst quality, low quality, extra digits, loli, loli face:1.2), low-res, blurry, text, logo, artist name, watermark + [LLM-generated negative prompt (only for Stable Diffusion)]