Taylor & Francis
Taylor & Francis Group

Journal of Engineering Design

ISSN: 0954-4828 (Print) 1466-1837 (Online) Journal homepage: www.tandfonline.com/journals/cjen20

Investigating intelligent generation of multimodal creative stimuli in conceptual design: strategies and implications

Zhuoshu Li, Pei Chen, Yexinrui Wu, Jiayi Yao, Hongbo Zhang, Xuanming Liu & Lingyun Sun

To cite this article: Zhuoshu Li, Pei Chen, Yexinrui Wu, Jiayi Yao, Hongbo Zhang, Xuanming Liu & Lingyun Sun (07 Jul 2025): Investigating intelligent generation of multimodal creative stimuli in conceptual design: strategies and implications, Journal of Engineering Design, DOI: 10.1080/09544828.2025.2527517

To link to this article: https://doi.org/10.1080/09544828.2025.2527517

Published online: 07 Jul 2025.

JOURNAL OF ENGINEERING DESIGN
https://doi.org/10.1080/09544828.2025.2527517

Taylor & Francis
Taylor & Francis Group

Investigating intelligent generation of multimodal creative stimuli in conceptual design: strategies and implications

Zhuoshu Li[a], Pei Chen[a,b], Yexinrui Wu[a], Jiayi Yao[c], Hongbo Zhang[a], Xuanming Liu[a] and Lingyun Sun[a,b]

[a]College of Computer Science and Technology, Zhejiang University, Hangzhou, People's Republic of China; [b]Zhejiang-Singapore Innovation and AI Joint Research Lab, Hangzhou, People's Republic of China; [c]School of Software Technology, Zhejiang University, Ningbo, People's Republic of China

ABSTRACT

Creative stimuli are pivotal sources of inspiration in conceptual design. As large generative models develop, AI has been increasingly utilised in creativity support tools to furnish designers with high-quality stimuli. However, the effects of stimuli vary across stages, rendering the certain types of creative stimuli generated by existing tools unadaptable to the dynamically changing conceptual design process. To explore the intelligent generation of multimodal creative stimuli in conceptual design, we first proposed an intelligent generation strategy to facilitate the autonomous generation of stimuli tailored to different stages. Subsequently, we designed and developed InsPilot, a system capable of proactively generating stimuli with consideration of factors including the design stage, design ideas, and the designer's input state. Through a comparative experiment with GPT-4V ($N = 12$), we examined the relative effectiveness of the system in supporting design creativity and delineated the relevant implications. Specifically, the creativity support provided by AI-generated stimuli include subdividing application scenarios and enhancing empathy in Rapid Divergence; refining, concretising, and transferring ideas in In-depth Divergence; identifying issues and focussing on core functions in Convergence, etc. The finding also revealed InsPilot's and GPT-4V's distinct impact on designers' behaviour and frequency of stimuli engagement.

ARTICLE HISTORY
Received 26 November 2024
Accepted 27 June 2025

KEYWORDS
Creative stimuli; generative AI; creativity support tool; conceptual design

1. Introduction

During the conceptual design process, creative stimuli serve as pivotal external sources of inspiration (Gomes et al. 2022), assisting designers in conceptualising design solutions that address specific problems and requirements (Benami and Jin 2008). These stimuli, typically in the form of texts and images, have the potential to help designers avoid issues such as insufficient inspiration and design fixation, while simultaneously enhancing the novelty (Goldschmidt and Sever 2011), practicality (Koronis et al. 2022), and diversity

CONTACT Pei Chen chenpei@zju.edu.cn College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, People's Republic of China, Zhejiang-Singapore Innovation and AI Joint Research Lab, Hangzhou310027, People's Republic of China

(Paay et al. 2023) of ideas. Designers can seek creative stimuli by exploring design communities like Pinterest[1], utilising card-based creative stimulation toolkits (Lomas, Karac, and Gielen 2021), and employing stimulus search tools such as DANE (Vattam et al. 2011). However, these conventional tools offer only a limited range of stimuli and fail to deliver highly targeted stimuli pertinent to specific design tasks.

With the advancement of large language models (LLMs) (Brown et al. 2020; P. Liu et al. 2023; Touvron et al. 2023) and text-to-image generation techniques (Podell et al. 2023; Rombach et al. 2022), Generative Artificial Intelligence (GenAI) has been able to rapidly create high-quality content across various modalities. This breakthrough has catalysed the integration of GenAI into creative stimulation tools for conceptual design (Fang et al. 2025; J. Jin et al. 2025). For instance, Idea Machine is adept at extending, revising, and combining designers' textual concepts to generate novel ideas (Di Fede et al. 2022), while BrainFax excels in generating visual stimuli based on designers' online dialogues (Verheijden and Funk 2023). These tools generate specific types of creative stimuli aligning with the design task, thus providing targeted support in the process of conceptual design.

However, the impact of creative stimuli on conceptual design is multifaceted and intricate, rendering the specific types of stimuli produced by current tools inadequate for the evolving conceptual design process. On one hand, conceptual design encompasses divergence and convergence stages, each requiring distinct types of stimuli. For example, J. Xu, Chao, and Fu (2020) found that words and phrases boosted creativity in the divergence stage but were less inspiring in convergence. Concurrently, Benton, Varotsis, and Vasalou (2019) highlighted the need for cautious consideration before integrating design examples into the design process. Nevertheless, most existing tools only generate stimuli of a single type and style, unable to cater to designers' varying needs at different stages. On the other hand, the initiative of generation also affects the effectiveness of creative stimuli, and proactively provided stimuli are generally more beneficial (Siangliulue et al. 2015). Despite this, most tools currently operate in a passive manner, relying on designers' prompts and unable to track the design process or generate stimuli in a proactive and timely manner aligning with the design stage.

To leverage AI-generated creative stimuli for augmenting designers' creativity, a critical question is how to intelligently generate multimodal creative stimuli throughout the conceptual design process. The term 'intelligently' implies two aspects: generating differentiated multimodal stimuli of appropriate content (e.g. relevant scenes, concepts, and products) for both divergence and convergence stages, and proactively generating creative stimuli tailored to the current design task and designer's status. Given the paucity of research providing practical experience or empirical evidence on this question, we aim to explore the strategy, impacts, and implications for the intelligent generation of stimuli through designing and evaluating a multimodal stimulus generation system.

The primary phases and outcomes of our research are illustrated in Figure 1. Initially, we identified appropriate types of stimuli for different design stages and developed prompt templates and an interaction pipeline to support autonomous generation without human intervention. Based on this, we implemented a system called InsPilot, capable of tracking designers' ideas, input states, and design stages, thereby generating tailored multimodal stimuli throughout the design process. Finally, we conducted a comparative study with GPT-4V to examine its impact on the conceptual design process through linkograph, thematic analysis, and behaviour coding.

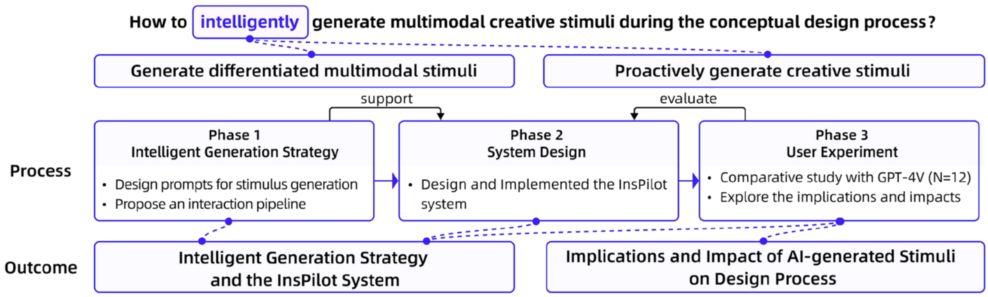

Figure 1. Overview of the research process.

The remainder of the paper is organised as follows: Section 2 reviews the relevant literature. Section 3 details the stimulus generation strategy, including the prompt design and interaction pipeline. Section 4 introduces the design and implementation of InsPilot. Section 5 describes the methodology of the user study. Sections 6 and 7 report the quantitative and qualitative findings, separately. Section 8 discusses implications, remaining challenges, and directions for future work. Section 9 concludes the study with a summary of key contributions.

2. Related work

2.1. Influence of creative stimuli on conceptual design

Creative stimuli, also termed inspiration or external stimuli, are crucial external sources of inspiration during the conceptual design process (L. A. Vasconcelos and Crilly 2016). They are essential not only for fostering ideas during the divergence stage but also for evaluating, selecting, and refining ideas in the convergence stage. Research indicates that the impact of creative stimuli on design is significantly affected by factors such as modality and content.

Existing studies have illuminated the varied effects of stimuli across different modalities. An experiment revealed that texts enhance knowledge recall while images promote active thinking (Shi et al. 2017). Another study utilised patents in different modalities as stimuli, revealing that detailed texts reduce idea quantity in the early stages but not in later stages (Chan et al. 2011). Ruiz-Pastor, Borgianni, and Chulvi (2024) further explored sustainability-oriented images' impact on the circularity and novelty of ideas. Compared to unimodal stimuli, the simultaneous presentation of textual and visual stimuli reduces textual ambiguity (Shi et al. 2017). Furthermore, Borgianni, Rotini, and Tomassini (2017) demonstrated that multimodal stimuli facilitated more creative ideas, which inspired us to utilise AI to generate both textual and visual stimuli in our following work.

As for the content of stimuli, existing work mostly adopts (1) abstract stimuli that provide descriptions of design scenarios, general concepts, or system attributes; and (2) concrete stimuli that include explicit functional examples or detailed design solutions (Ezzat et al. 2020). Studies have proven that abstract stimuli foster abstract thinking (Huang, Gino, and Galinsky 2015), thereby effectively increasing idea quantity (L. A. Vasconcelos and Crilly 2016) and novelty (Gonalves, Cardoso, and Badke-Schaub 2012; Paay et al. 2023). While another study found that both categorical and specific examples facilitated originality (Yuan et al. 2023), an experiment revealed that concrete examples increased design fixation and reduced idea quantity (Ezzat et al. 2020).

2.2. Traditional creative stimulation tools

In the divergence stage of conceptual design, the Internet and search engines are the main sources of stimuli, helping designers maintain the frequency of idea generation and avoid being stuck. However, designers occasionally lack focus during the search process (Herring et al. 2009), which has promoted the development of categorisation-based (Yee et al. 2003) or image-based (Fogarty et al. 2008) search engines. Despite that, design fixation often confines designers to search within limited keywords, impeding their access to a diverse range of stimuli and subsequently diminishing the diversity of ideas.

To foster diverse ideas, previous research extracted concepts from materials including patents, papers, and products, presenting them in the form of creative stimulation cards (Golembewski and Selby 2010; L. Vasconcelos et al. 2024; Mora, Gianni, and Divitini 2017; Yilmaz et al. 2016). For example, X. Jin et al. (2021) extracted 40 design heuristics from thousands of patents to assist in the design of intelligent canteens. Other similar card toolkits, such as Card Mapper (Darzentas et al. 2019) and Design Space Cards (Lomas, Karac, and Gielen 2021), provide multimodal stimuli for different fields. However, the fixed content of card toolkits limits their applicability to various design tasks, necessitating that designers manually select cards to obtain truly inspiring stimuli, thus increasing cognitive load. To address this limitation, interactive systems were developed to support stimulus provision (Chakrabarti et al. 2005; Vandevenne et al. 2016). For instance, Althuizen and Wierenga (2014) developed a system featuring a vast array of design cases, supporting adjustment of semantic distance between cases and tasks. DANE provided an interactive environment for exploring structures, behaviours, and functions from bionic databases (Vattam et al. 2011). Despite their interactive and flexible nature, these systems are confined to existing databases, limiting their adaptability to various design tasks.

While creative stimuli have found widespread application in the divergence stage of conceptual design, their utilisation in the convergence stage is less prevalent although several studies have validated stimuli's effect on convergent thinking (Sassenberg et al. 2017). For example, the study by W. Liu et al. (2022) employed images and sounds of the natural environment as stimuli to improve participants' sense of pleasure in completing convergence tasks.

The above work demonstrates the significant value of multimodal stimuli in fostering conceptual design, but the fixed or limited content of stimuli cannot fully meet designers' diverse needs. This highlights a pronounced demand for tools able to dynamically provide creative stimuli based on the design task and the designer's ideas, and GenAI serves as a promising solution to the demand.

2.3. AI-based creativity support tools

The recent advancement of AI has triggered numerous AI-based search tools and generative tools. AI's capabilities of information comprehension and processing empower search tools to provide more tailored results (Van Noorden and Perkel 2023). Systems including IdeateRelate (X. T. Xu et al. 2021) and OntoTag (2010) supported semantic-based retrieval for texts and images. Several tools even accepted designers' sketches as input (J. Han et al. 2018; J. Kim, Maher, and Siddiqui 2021; Lindley et al. 2013), with some supporting adjustments to the search results based on similarity (Karimi et al. 2020) and semantic distance (J. Kim and Maher 2023; Kang et al. 2021; Q. Chen et al. 2023). However, AI-based search tools can easily lead to designers' deviation from the original idea during the searching process (Mozaffari et al. 2022), and stimuli of similar styles provided by some search tools may lead to design fixation (Herring et al. 2009).

By incorporating GenAI, such as ChatGPT,[2] Midjourney[3], Stable Diffusion,[4] and DALL·E,[5] many CSTs can now transform vague descriptions or rough sketches into concrete design proposals (D. Wu et al. 2023; Zhang et al. 2024). For example, PromptMagician supported specifying personalised criteria for creating images (Feng et al. 2023), PromptPaint built a paint medium-like interface for iterative refinement of the generated image effects (Chung and Adar 2023). GPSdesign (P. Chen, Wu, et al. 2025c) and CoExploreDS (P. Chen, Yao, et al. 2025d) structure the design process through a problem–solution co-evolution model, supporting designers in exploring design spaces systematically. FusionProtor (Zhang et al. 2025) facilitates rapid ideation by transitioning physical prototypes into virtual, high-fidelity representations and supports component-level iteration and simulation. These works demonstrate GenAI's advantages of assisting design, but the early presentation of specific design cases might deepen design fixation (Ezzat et al. 2020). In addition, Urban Davis et al. (2021) addressed that conceptual design often involves a divergent ideation process, as opposed to generating several candidate design solutions at the outset.

Compared to directly generating design solutions, generating differentiated stimuli is more promising for mitigating design fixation and fostering creativity. Taking advantage of LLMs, Idea Machine could extend, rewrite, and merge designers' ideas, proposing new ideation directions based on existing design proposals (Di Fede et al. 2022). AskNatureGPT proposed the stimuli generation method to automatically search for problems and transfer biological analogy (L. Chen, Cai, et al. 2025). I-card integrated GenAI to provide applicable design methods, design knowledge and interactive support (L. Chen, Cheang, et al. 2025). PopBlends generated numerous words based on designers' keyword input to facilitate continuous divergence (S. Wang et al. 2023), and Yun et al. (2022) fine-tuned a GPT with data from the Red Dot design award winners to generate high-quality stimuli. Similarly, text-to-image generation models are widely employed to generate and blend visual stimuli (Sun et al. 2025). It included the style generation (M. Zhou et al. 2024), colour-concept association (Hou et al. 2025), sketch-based generation and edition (Lin et al. 2025; Peng, Koch, and Mackay 2025), and user interface generation (Mozaffari et al. 2022). Several tools even supported multimodal stimulus generation for greater effectiveness in promoting creativity (Borgianni, Rotini, and Tomassini 2017). For instance, StepIdeator (Yao et al. 2025) employed GenAI to support designers in externalising concepts with seamless transition and refinement of mixed-design representations including design briefs, sketches, model images, and renderings.

Unlike CSTs for directly generating the final proposals, a common feature of stimulus generation tools is to preset the types or styles of the stimuli, thereby tailoring the stimuli to the current design task. While enlightened by this feature, we also found several areas to improve existing work. Firstly, existing tools can only generate specific types of stimuli (e.g. figures in the mood board, product description texts) for a particular step of the divergence stage. However, designers require diverse stimuli in all design stages, with multimodal stimuli commonly bringing more inspiration (Borgianni, Rotini, and Tomassini 2017; Nijstad, Stroebe, and Lodewijkx 2002). Secondly, while proactively provided stimuli have been proven more effective (C. Zhou, Zhang, and Yu 2023; Siangliulue et al. 2015), existing tools can only generate stimuli passively based on designers' instructions, unable to track the design process and dynamically adjust to designers' evolving needs and contexts. Therefore, current stimulus generation tools could be advanced by involving diverse preset stimuli and incorporating a reasonable generation pipeline, which would be explored in our work.

3. Intelligent stimulus generation strategy

This section explores the utilisation of multimodal large language models (MLLMs) to produce textual or visual stimuli for designers proactively.

3.1. Determining creative stimuli for different design stages

To determine appropriate stimulus types for our generation strategy, we first define the stages of conceptual design. Conceptual design refers to the process of translating unarticulated needs into concrete product functionalities (Andreasen, Hansen, and Cash 2015). A widely adopted framework is the classical Double Diamond Model (Design Council 2005), which outlines four phases: Discover, Define, Develop, and Deliver. Our generation strategy specifically targets the latter two phases, Develop and Deliver, which are dedicated to design ideation.

In the scope of this paper, since we focus on the rapid generation and iteration of ideas, we mainly adopt sketch-based visualisations instead of virtual presentation or full-scale prototypes (Kuys, Ranscombe, and Zhang 2023). To investigate with greater granularity, following the work of P. Chen et al. (2024), we further refine the Develop and Deliver phases into the three stages in Figure 2 and elaborate the adopted design representations referring to Pei, Campbell, and Evans (2011) as follows:

- Rapid Divergence: Designers rapidly generate ideas based on the design task in forms of texts and sketches;
- In-depth Divergence: Designers selectively refine solutions from the previous phase and continue generating ideas. In addition to sketches of products, designers might also sketch usage scenario and technical illustrations to explore the concept's context and feasibility;
- Convergence: Designers evaluate existing ideas and formed a final sketch-based solution with brief illustrations including usage scenario and diagrams of product structures and interaction process.

It is noteworthy that although the process was presented as linear for clarity of the generation strategy, the In-depth Divergence Stage includes iterative ideation and refinement.

To explore diverse stimulus combinations' influence on creativity across different stages, P. Chen et al. (2024) conducted a 3 × 3 between-subject study involving textual and

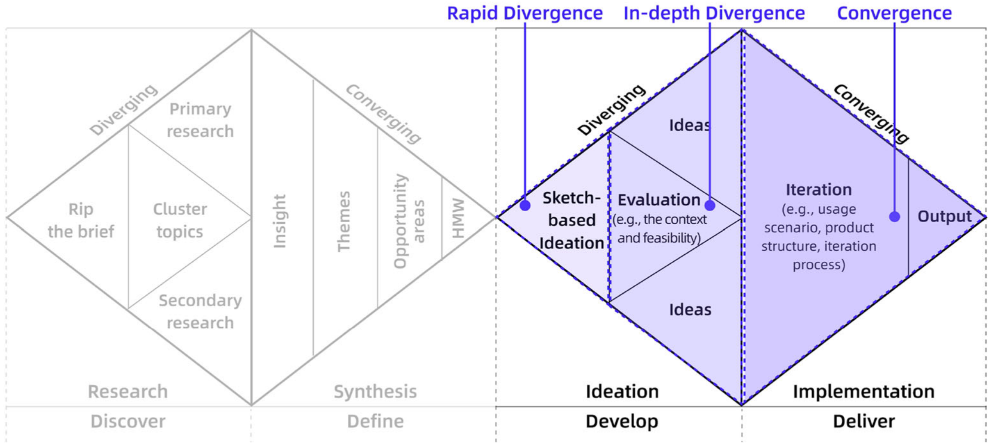

Figure 2. Overview of the Double Diamond Model and the three corresponding design stages in the generation strategy.

visual stimuli with different fidelities (including concept/product texts and scene/product images). Based on their findings, the combinations for the following prompt design and interaction pipeline were selected with data support, including: (1) concept texts and scene images for Rapid Divergence, which provide a pathway for designers to recall past experiences through scene images and then conceive possible product forms with concept texts; (2) product texts and scene images for In-depth Divergence, in which the scenes help designers recall ignored issues in their ideas and the product description assists in idea refinement; and (3) product texts with both types of images for the Convergence Stage, because simultaneous presentation of product text and images helps designers detail a product's form, appearance, structure, etc., and scene images assist them in selecting more practical solutions.

3.2. Prompt design

Prompts denote the textual content utilised to express users' requirements to GenAI, which decide the content and quality of the generated result. We designed and iterated prompt strictly following the guidelines proposed by Zhao et al. (2023), which suggested providing explicit and specific instructions including limiting symbols, output format requirements, and examples.

The iteration process of prompt included coarse-grained and fine-grained stages. Initially, we developed evaluation criteria for different stimuli, including instruction compliance, content relevance, output stability, definition compliance, diversity, and rationality (details in Appendix A.1). Each co-author then independently decomposed the generation process into minor steps (T. Wu et al. 2022) and refined the prompt until he/she believed the generated stimuli met these criteria. Subsequently, we invited five researchers specialising in GenAI to review the generated stimuli based on these criteria and select the optimal prompt for fine-grained iteration. Then, researchers and co-authors collaboratively reviewed the selected prompt, experimenting with different structures and wordings to reduce output uncertainty, until everyone agreed that the stimuli met the evaluation criteria.

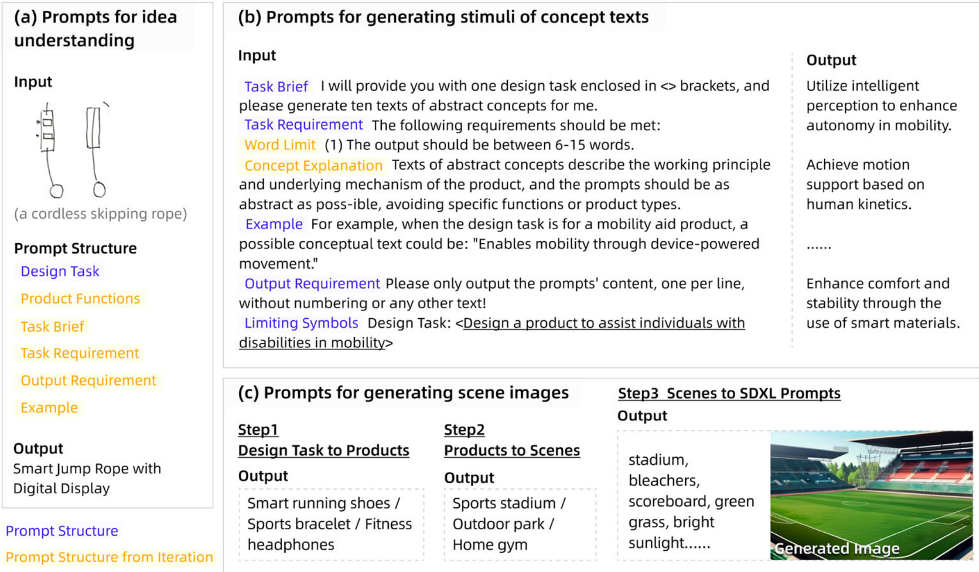

Figure 3. Prompts used for (a) idea understanding and (b) generating stimuli of concept texts, and (c) process of generating scene images.

3.2.1. Prompts for idea understanding

The initial step employs an MLLM to interpret and transcribe designers' ideas, conveyed through texts and sketches, into a textual description. To achieve this, we adopted GPT-4V, an MLLM developed by OpenAI, capable of understanding images and generating corresponding descriptions. After interactive experiments, we observed improved understanding by GPT-4V when the prompts incorporated both the design task and product functions (see Figure 3(a)), and GPT-4V's output was then used as input for generating further texts and images. Details of the full prompt are available in Appendix A.2.

3.2.2. Prompts for generating textual stimuli

When generating concept texts, initial attempts using GPT with example-based prompts did not meet the evaluation criterion of definition compliance. To refine this, we clarified the definition of 'concept texts' within the prompts to avoid misunderstanding, and limited the output length to avoid verbose text containing unnecessary details. The final prompt for generating concise concept texts is shown in Figure 3(b). For product text generation, GPT's initial outputs were relatively brief and lacked details. To address this, we integrated quantitative requirements into the prompts to mandate the inclusion of at least three functionalities in GPT's outputs (details in Appendix A.2).

3.2.3. Prompts for generating visual stimuli

Stable Diffusion XL (SDXL) (Podell et al. 2023) was adopted to generate visual stimuli due to its proficiency in producing diverse high-quality images from textual prompts. Although GPT-4V also supports image generation, its restrictions on the number of images generated per minute fail to satisfy designers' extensive demand for stimuli, leading to our preference for SDXL. The image generation strategy comprises two parts. First, to assure image quality, we refined the content and weighting of SDXL prompts multiple times and culminated in effective ones (see Appendix A.2). Subsequently, we iterated GPT prompts, directing GPT to generate adjustable prompts reflective of design tasks and designers' ideas. We segmented the generation process into three steps: generating relevant products from the design task, deriving application scenarios from these products, and formulating prompts for scene images based on these scenarios. This approach yielded scene images aligning with our requirements (see Figure 3(c)). Similarly, for product images, we split the generation of adjustable prompts into two stages: generating product ideas from the task and creating SD prompts based on these products (more details in Appendix A.2).

3.3. Interaction pipeline

In the actual conceptual design process, creative stimuli do not appear in isolation, highlighting the necessity to integrate the aforementioned prompts into a seamless interaction pipeline. With primary consideration of the following two aspects, the final established interaction pipeline is illustrated in Figure 4 as a part of the intelligent generation strategy:

- Ensure coherence among different types of stimuli. Simultaneously appearing stimuli should have a connection to avoid confusion for designers.
- Reduce repetitive generation steps. When generating different stimuli, there are repetitive steps that may result in more rounds of generation and a longer generation time. The interaction pipeline should integrate repetitive generation steps to enhance the efficiency of generation.

During the Rapid Divergence Stage, the design task serves as input for GPT-4, and therefore the generated stimuli are parallel to ideas produced at this stage. In the other two stages, GPT-4V first comprehends the designer's ideas before they are input into GPT-4. For consistency and efficiency of the generation process, for all stages, GPT-4 should initially generate relevant products based on the input. Then, GPT-4 and SDXL generate texts and images separately using prompts obtained in Section 3.2.

The generated stimuli can be presented to the designer either proactively or passively. On one hand, we monitor the designer's input status and automatically display creative stimuli when no input is detected in a period (i.e. 15 seconds in the Rapid Divergence Stage, 12 seconds in the other two stages). The time duration is determined based on the designers' behaviour in the formative study. Meanwhile, designers can also request to view creative stimuli proactively without limitation.

Based on the above generation process and interaction logic, this interaction pipeline can guide generative models to intelligently generate creative stimuli based on the design task, design stage, and the designer's status.

4. System design

This section introduces the design of InsPilot (short for Inspiration Pilot), following the prompt and interaction pipeline obtained in Section 3. The core objective of InsPilot is to intelligently generate inspirational stimuli based on the designer's design tasks, design status, and the current design stages, assisting designers in gaining afflatus and enhancing their creativity.

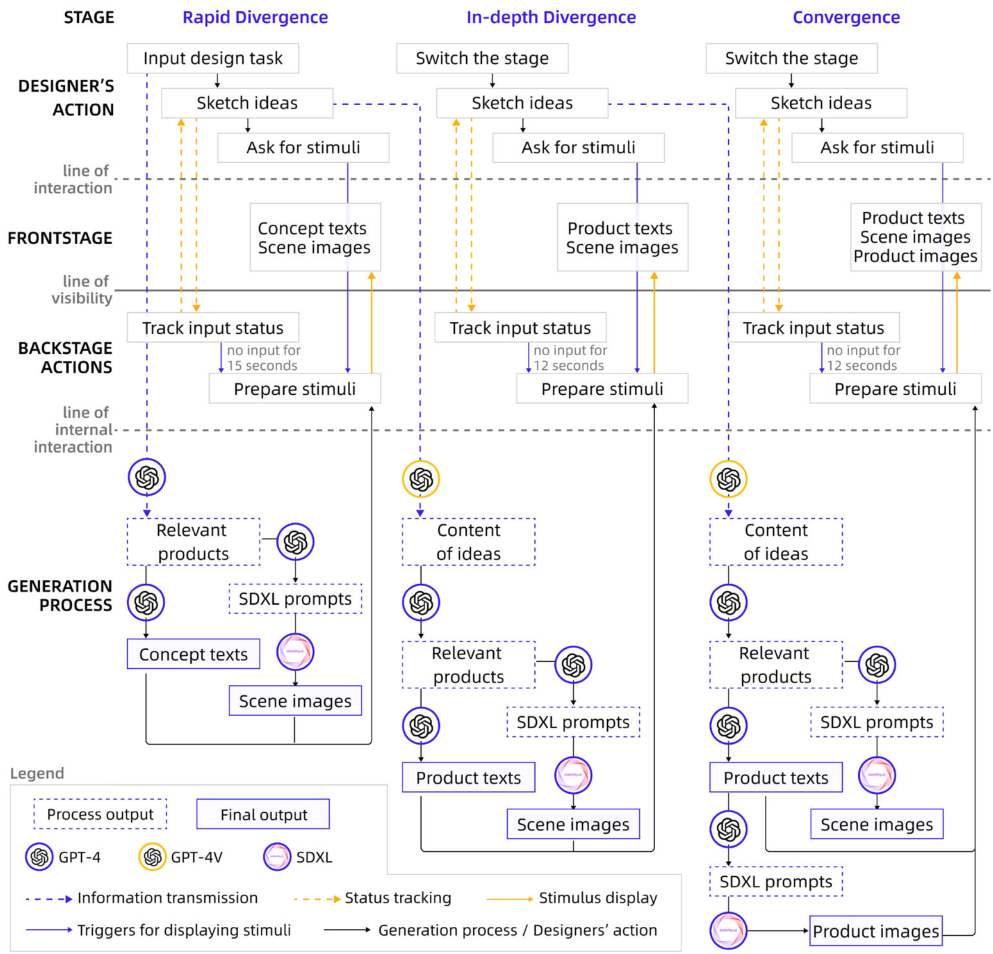

Figure 4. Interaction pipeline for intelligently generating creative stimuli.

4.1. Interaction design

As shown in Figure 5, InsPilot includes three modules: Basic Information Module, Design Idea Module, and Creative Stimuli Module.

The Basic Information Module contains the following functions (see Figure 5(a–c)): (1) User information input: The system will create a new document for new users or retrieve and display the historical design ideas for existing users. (2) Design task input: The designer needs to input the design task for InsPilot to generate stimuli relevant to the task. (3) Idea selection: The designer can choose multiple concepts in the Convergence Stage for InsPilot to generate creative stimuli based on the selected ideas.

The Design Idea Module is where designers formulate their design proposals (see Figure 5(d–g)), consisting of the following parts: (1) Stage selection: The designer can flexibly switch between the three stages, and the current stage will affect the types of generated stimuli. (2) Sketching toolbox: InsPilot offers necessary tools including brush, text box, eraser, undo, clear, and save. The text box tool supports keyboard and voice input of text

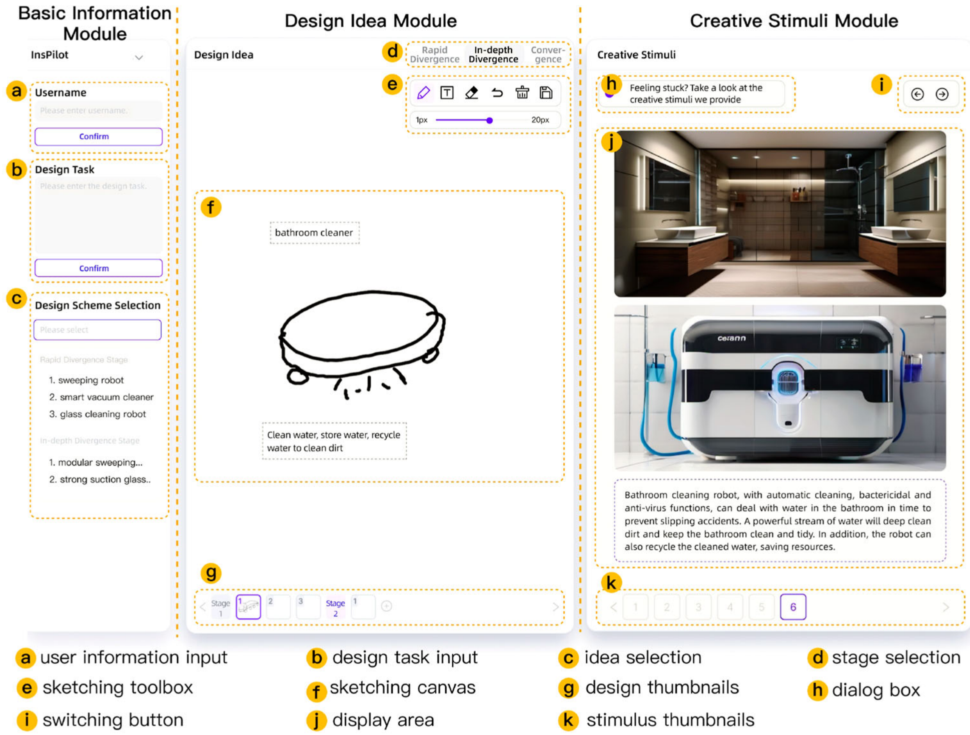

Figure 5. Interface and main modules of InsPilot.

anywhere on the canvas. (3) Sketching canvas: The designer can sketch the ideas on the canvas, as illustrated in Figure 5(f). (4) Design thumbnails: The designer can click on the thumbnail to view existing ideas.

The Creative Stimuli Module comprises four parts (see Figure 5(h–k)): (1) Dialog box: The system provides concise explanations for the current stimulus, with messages like 'I have generated a creative stimulus for you based on the first idea in the Rapid Divergence Stage'. (2) Switching button: The designer can view automatically generated stimuli or proactively generate new stimuli by clicking the button. (3) Display area: This area displays multimodal stimuli for the designer following the interaction pipeline. In this area, images appear first, and texts only appear after the designer clicks the switching button. In this way, we aim to address the observed tendency of designers to focus more on the text and neglect visual information when both modalities of stimuli are presented simultaneously as suggested by P. Chen et al. (2024). (4) Stimulus thumbnails: Previously displayed stimuli are shown as thumbnails below the display area, allowing designers to switch by clicking.

To more intuitively compare the differences between the interaction pipeline of InsPilot and GPT-4V, we summarise their respective features in Table 1.

4.2. Implementation

InsPilot is a web-based system that utilises React as its frontend framework, with Ant Design components[6] for styling. The function of sketching is realised using a freehand vector

Table 1. Comparison on interaction-/design-related features of InsPilot and GPT-4V.

Feature	Sub-feature	InsPilot	GPT-4V
Interaction-related features	Input modality	Text, sketch, speech	Text, image, speech
	Stimulus generation	Based on design task, the designer's existing ideas, and preset effective prompt	Fully based on the designer's prompt and image input
	Stimulus presentation	Proactively present or when requested by the designer	Only when requested by the designer
	Designer's control over the generated content	Partial control (over the direction of generation)	Full control without restriction or guidance
Design-related features	Design stages	Three conceptual design stages, allowing free switching	Not clearly distinguished
	Stimulus modality	Text and image, tailored to different design stages	Text or image, as required by the designer
	Stimulus content	Concept/product text and scene/product image, tailored to the stage	Fully depends on the designer's requirement

drawing component called 'react-sketch-canvas'. The backend of InsPilot, developed in Python with Flask, leverages MongoDB for storing user information and design ideas. We utilise GPT-4 to generate textual stimuli and use GPT-4V to understand design ideas via OpenAI's API. SDXL is deployed locally to generate image stimuli, and six NVIDIA GTX 3090 GPUs are used for the parallel generation of images.

5. User experiment

The experiment primarily serves two goals: (1) to validate the feasibility of the proposed interaction pipeline and preliminarily evaluate InsPilot's effectiveness in providing creativity support for conceptual design, and (2) to explore the impacts and implications of creative stimulus generation, providing a reference for future research on utilising AI-generated stimuli in conceptual design.

5.1. Experiment design

5.1.1. Conditions

To mitigate the impact caused by designers' various prior experiences with AI-based CSTs and their prompt skills, we employed a within-subject design. For experimental condition, participants designed with InsPilot using a 12.9-inch tablet and stylus. For control condition, they sketched ideas in InsPilot using the same tablet and acquired stimuli from GPT-4V on another tablet.

We chose GPT-4V for the control condition primarily because: (1) GPT-4V is among the most effective MLLMs currently available, with robust capabilities in generating texts and images, aligning with the modalities of stimuli in our system. (2) InsPilot was implemented with the API of GPT-4V, and therefore, the quality of generated stimuli from our system and GPT-4V are comparable.

In the experiment, participants performed two different conceptual design tasks under both conditions. Task 1 was to design a product to assist senior citizens in their mobility, and task 2 was to design a smart household cleaning product (detailed descriptions in Appendix 2). We systematically counterbalanced the sequence of conditions and tasks to avoid order effects.

5.1.2. Participants

This experiment involved 12 participants recruited from social media (hereinafter referred to as P1–P12), aged between 21 and 25 ($M = 22.42$, $SD = 1.19$), including 5 males and 7 females. All participants were graduate or senior undergraduate students from design-related majors, with over two years of experience in conceptual design. They demonstrated proficient deliberation and expression skills, capable of articulating design proposals through texts and sketches. All participants received compensation after the experiment.

5.1.3. Procedure

Before the experiment, participants signed an informed consent form and provided basic information including their age, gender, major, grade, design courses they have taken, and experience with GenAI. Then, each participant underwent a five-minute practice on thinking aloud.

The experiment consisted of two sessions. In each session, we introduced the system's primary functions and operations to participants. Subsequently, participants were given 32 minutes to complete the design task while thinking aloud, going through the stages of Rapid Divergence, In-depth Divergence, and Convergence. Participants were encouraged to illustrate their ideas using sketches and descriptive texts. Considering the experiment's feasibility and referring to other similar work (Hatcher et al. 2018; Paay et al. 2023), the 32-minute experiment setup is more like a brainstorming session covering main design stages rather than a complete and actual design process. However, to simulate a real design process, no time restrictions were imposed on each stage, allowing participants to freely allocate time or switch between stages. The tablet screen was recorded throughout the experiment.

After each session, participants filled out questionnaires (see Section 5.2) and received a semi-structured interview. The interview primarily focussed on participants' feedback on the used system, including inquiries about the reasons behind their ratings and the reasons for changes in scores at different stages. Participants were also asked about differences between the design process supported by AI-generated stimuli and their previous design process, as well as the strengths and weaknesses of InsPilot and GPT-4V. The interview also included questions such as: How do you use AI in your daily conceptual design process, and how does it assist you in completing design tasks? Are there instances where you seek assistance from AI but find it insufficient to provide such assistance? The experimenter adapted questions flexibly based on participants' responses.

5.2. Evaluation metrics

Participants evaluated two systems quantitatively in terms of the System Usability Scale (SUS) (Bangor, Kortum, and Miller 2008), the Creativity Support Index (CSI) (Cherry and Latulipe 2014), and the creativity of their design output.

For output creativity, participants' evaluation criteria included novelty, practicality, and diversity. In addition, three experts were invited to the assessment. All experts possessed

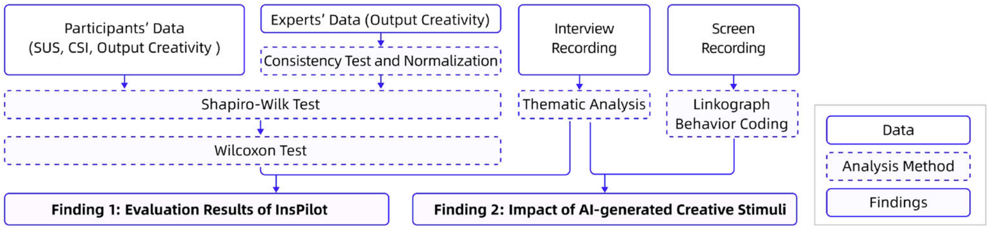

Figure 6. Method of data analysis.

over six years of experience in conceptual design and were independent and specialised in evaluating design output.

Experts' evaluation criteria were the same as participants but with a slightly different scoring standard. Referring to the work of Sarkar and Chakrabarti (2011), novelty and diversity were still evaluated using a seven-point scale, while practicality was scored based on the product of three sub-criteria: (1) Level of importance: a five-point scale, where 1 = not important at all, 5 = extremely important. (2) Rate of popularity: a range from 0–1, where 0 = no one would use it, 1 = everyone would use it, rounded to one decimal place. (3) Rate of use: a range from 0–1, where 0 = no one would use it, 1 = everyone would use it, rounded to one decimal place.

5.3. Data analysis

Data analysis follows the workflow in Figure 6. We first applied the Shapiro-Wilk test to assess the normality of participants' data. All *p*-values were below 0.05, indicating non-normal distributions; thus, the Wilcoxon signed-rank test was used for further comparison. For expert ratings, Kendall's W test confirmed strong inter-rater agreement across all stages ($W \geq 0.6$, $p < 0.001$). To account for individual scoring ranges, ratings were linearly normalised using $X_{normalized} = (X_i - X_{\min})/(X_{\max} - X_{\min})$, where X_i represented the expert rating and $X_{normalized}$ presented the normalised rating. The normalised data were then analysed following the same method as participants' data.

For qualitative results, we first utilised linkograph to clarify the connection between the generated stimuli and participants' ideas. The link types were determined based on prior work (Hatcher et al. 2018), detailed in Section 7.1. We also conducted a thematic analysis of the interview to explain the effectiveness of InsPilot and support further discussion. Based on the transcribed texts from the interview recordings, two co-authors independently identified the strengths and weaknesses of InsPilot and GPT-4V at each stage and throughout the entire design process and then synthesised them into summary codes. Subsequently, the two co-authors discussed their coding results together and addressed any discrepancies in the coding until a consensus was reached.

In addition, we employed behaviour coding to extract insights from the screen recording, primarily focussing on (1) participants' time distribution across three stages, (2) the timestamp at which participants viewed stimuli or communicated with GPT, and (3) the duration of conceiving each idea. The coding results helped us further understand how the differences between InsPilot and GPT-4V impact designers' behaviour.

6. Finding 1: evaluation results of InsPilot

This section presents the evaluation results of InsPilot with comparison to GPT-4V. The quantitative results will be introduced while quoting participants' qualitative feedback, with P1 denoting the first participant.

6.1. System usability scale

According to SUS's standard, both the average usability scores for InsPilot ($M = 86.46$, $SD = 7.39$) and GPT-4V ($M = 76.25$, $SD = 7.60$) surpassed the acceptance threshold (i.e. 70). As indicated by the Wilcoxon test, the overall usability of InsPilot was significantly higher than that of GPT-4V ($p = .005$), illustrating the high usability of InsPilot.

6.2. Creativity support index

The CSI of InsPilot ($M = 85.58, SD = 9.05$) was significantly higher than GPT-4V ($M = 74.83$, $SD = 9.76$) with a p-value of 0.002, indicating that InsPilot provided greater creativity support to designers during the design process. Specifically, InsPilot outperformed GPT-4V in all dimensions, with notable advantages in the Exploration, Expressiveness, and Immersion dimensions.

6.3. Output creativity

The output creativity of InsPilot and GPT-4V is compared based on the quantity of ideas, participants' ratings, and experts' ratings. For the quantity of ideas, the Wilcoxon test indicated a significant difference between InsPilot and GPT-4V in both the Rapid Divergence Stage ($p = .011, M_{InsPilot} = 4.92, SD_{InsPilot} = 0.95, M_{GPT-4V} = 3.75, SD_{GPT-4V} = 1.36$) and In-depth Divergence Stage ($p = .014, M_{InsPilot} = 2.58, SD_{InsPilot} = 0.86, M_{GPT-4V} = 2$, $SD_{GPT-4V} = 0.58$). The results of participants' and experts' ratings are shown in Table 2.

Participants' Ratings. In the Rapid Divergence stage, idea diversity was significantly higher with InsPilot than GPT-4V ($p = .006$). Participants noted that creative stimuli from both systems supported the generation of novel and practical ideas. In the In-depth Divergence stage, InsPilot led to significantly higher idea novelty compared to GPT-4V ($p = .034$). Participants attributed this to the product texts provided by InsPilot, which helped them envision functions beyond those of existing commercial products (P1). In the Convergence stage, ideas generated with InsPilot were rated significantly higher in practicality ($p = .014$). For instance, P4 described how the scene images enabled them to 'roughly envision real-world application scenarios', which in turn supported the refinement of usage workflows and enhanced idea feasibility.

Experts' Ratings. In the Rapid Divergence Stage, the novelty and diversity of ideas produced with InsPilot were significantly higher than GPT-4V ($p_{novelty} = .046, p_{diversity} = .049$). Participants using InsPilot significantly outperformed GPT-4V in terms of diversity in In-depth Divergence ($p = .027$) and novelty in Convergence ($p = .047$). Overall, the significance and average scores of experts' ratings indicate InsPilot's positive influence on the novelty and diversity of ideas in all stages, while the differences in practicality between InsPilot and GPT-4V were not obvious.

Table 2. Comparisons on participants' and experts' ratings of output creativity. Experts' ratings were normalised to 0–1 before analysis.

		Novelty					Practicality					Diversity				
		InsPilot		GPT-4V			InsPilot		GPT-4V			InsPilot		GPT-4V		
Stage	Scorer	M	SD	M	SD	*p* value	M	SD	M	SD	*p* value	M	SD	M	SD	*p* value
Rapid Divergence	Participant	5.25	1.16	4.83	1.14	0.257	5.75	0.83	5.42	0.64	0.102	5.58	1.04	4.42	0.95	**0.006***
	Expert	0.53	0.12	0.40	0.15	**0.046***	0.31	0.05	0.30	0.06	0.699	0.55	0.25	0.41	0.30	**0.049***
In-depth Divergence	Participant	5.42	0.76	4.75	0.83	**0.034***	6.00	0.71	5.58	0.86	0.102	5.00	1.15	4.42	0.86	0.239
	Expert	0.62	0.16	0.54	0.19	0.416	0.31	0.11	0.35	0.08	0.248	0.50	0.27	0.27	0.19	**0.027***
Convergence	Participant	5.83	1.07	5.50	1.12	0.317	6.42	0.64	5.83	0.80	**0.014***			/		
	Expert	0.74	0.20	0.52	0.26	**0.047***	0.37	0.13	0.39	0.21	0.780					

		Parallel link	Incremental link	New idea link	Alternative link	Tangential link
Rapid Divergence	InsPilot	0.00	1.75	3.25	0.42	1.58
	GPT-4V	0.67	1.50	0.25	0.67	0.17
In-depth Divergence	InsPilot	0.08	2.83	1.33	1.75	0.33
	GPT-4V	0.33	1.08	0.08	0.42	0.00
Convergence	InsPilot	0.00	1.58	0.33	0.67	0.00
	GPT-4V	0.00	0.75	0.00	0.42	0.00

Figure 7. Average numbers of links between stimuli and ideas in InsPilot and GPT-4V at different stages.

7. Finding 2: impact of AI-generated creative stimuli on conceptual design process

7.1. Links between stimuli and ideas

The average numbers of different types of links in InsPilot and GPT-4V and examples of stimuli are presented in Figures 7 and 8, including (1) *P*: parallel links that directly repeat the stimulus and only modify the wording (e.g. Figure 8(b)), (2) *I*: incremental links that add functions or refine the existing stimulus (e.g. Figure 8(e)), (3) *N*: new idea links that obtain a completely new solution from the problems of existing scenarios or ideas in the stimulus (e.g. Figure 8(a)), (4) *A*: alternative links that modify an element of the stimulus or apply it to a new scenario (e.g. Figure 8(d)), and (5) *T*: tangential links that produce ideas cognitively associated with the stimulus, such as having the same working principle or mechanism (e.g. Figure 8(c)).

Analyzing the number and type of links could help to uncover designers' different ideation patterns in InsPilot and GPT-4V. For example, in Rapid Divergence, when using InsPilot, participants mostly conceived ideas by identifying potential issues from the scene images (*N*), supplementing functions to the concept texts (*I*), and finding associations between the concept texts and other products related to the scene images (*T*). In GPT-4V, participants mostly built incremental links with occasional parallel links and alternative links. According to prior work, more parallel links in GPT-4V indicate more repetition and the lack of innovation, while more tangential links in InsPilot represent the generation of more novel ideas in divergence stages (Hatcher et al. 2018). Similarly, during In-depth Divergence, incremental links were mostly observed in both InsPilot and GPT-4V, which indicated that participants intensively utilised details from the generated stimuli to refine their existing ideas, while participants formed relatively more new idea links with InsPilot.

7.2. Creativity support from AI-generated stimuli

This section synthesises how creative stimuli support conceptual design across the three stages (see Figure 9) and discusses the shortcomings of both InsPilot and GPT-4V. Texts in bold are the codes from thematic analysis.

Rapid Divergence Stage. In Rapid Divergence, textual stimuli from both InsPilot and GPT-4V provided direction for ideation. On one hand, the generated texts helped designers explore different ideation directions in a systematic way. For instance, InsPilot generated

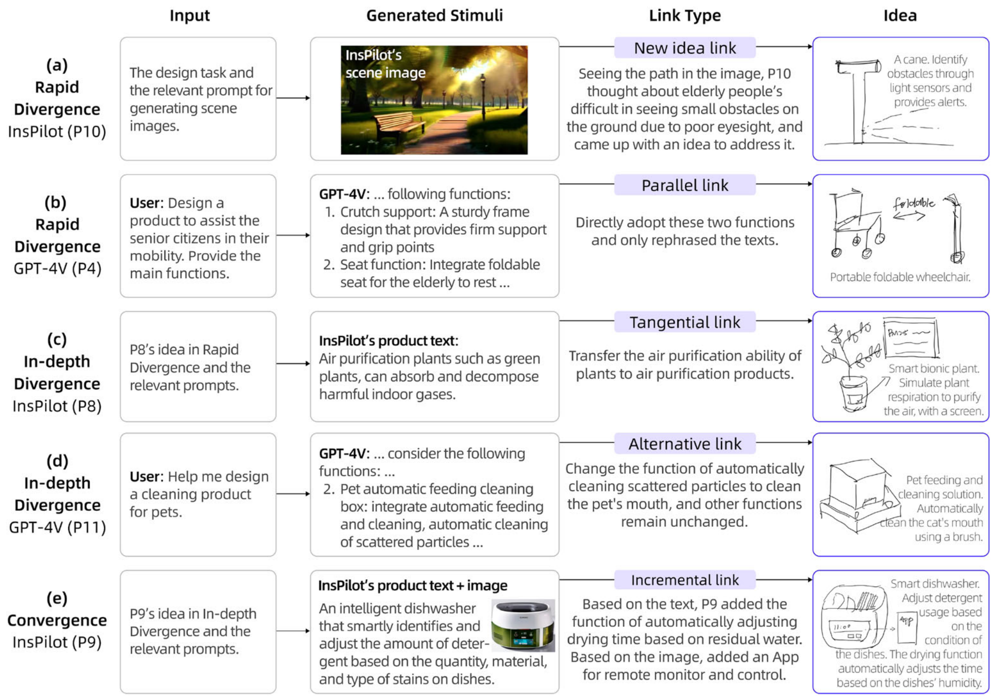

Figure 8. Typical examples of the links between stimuli and ideas.

abstract concepts based on the task 'design an assistance product for senior citizens' mobility', including 'sensing the environment to aid walking' and 'confirming health status through information exchange', helping participants 'know most of the possible directions for ideation' (P2). On the other, aligning with propulsion theory (Sternberg 2003), the textual stimuli from InsPilot and GPT-4V helped further expand the ideation directions, facilitate designers' redefinition of design solutions, and 'brought enlightenment during creative blocks' (P6).

In addition, InsPilot's scene images helped participants subdivide the product's application scenarios. When designing household cleaning products, scene images helped participants 'narrow a whole family scene down to specific spaces, like the kitchen and bathroom, making me diverge more logically' (P5). Furthermore, the scene images expanded application scenarios, redirecting designers to underexplored scenarios (Sternberg 2005). Another role of scene images was to enhance designers' empathy, helping them recall the needs and pain points of the scenarios. In this way, designers not only perceived stimuli, but also actively found meanings from the images (Yamamoto and Nakakoji 2005). For example, 'the forest path in the picture brought me into that scene. I would imagine the difficulties elderly people might have when walking in this environment' (P10).

Despite that, the absence of product images at this stage hindered two participants from visualising the product's appearance. Regarding GPT-4V, six participants noted the generation of irrelevant or repetitive content. Furthermore, three felt the generated content by GPT-4V was overly broad and macroscopical, detracting their focus from specific design problems. For example, when P3 'asked GPT to design household cleaning products, it only

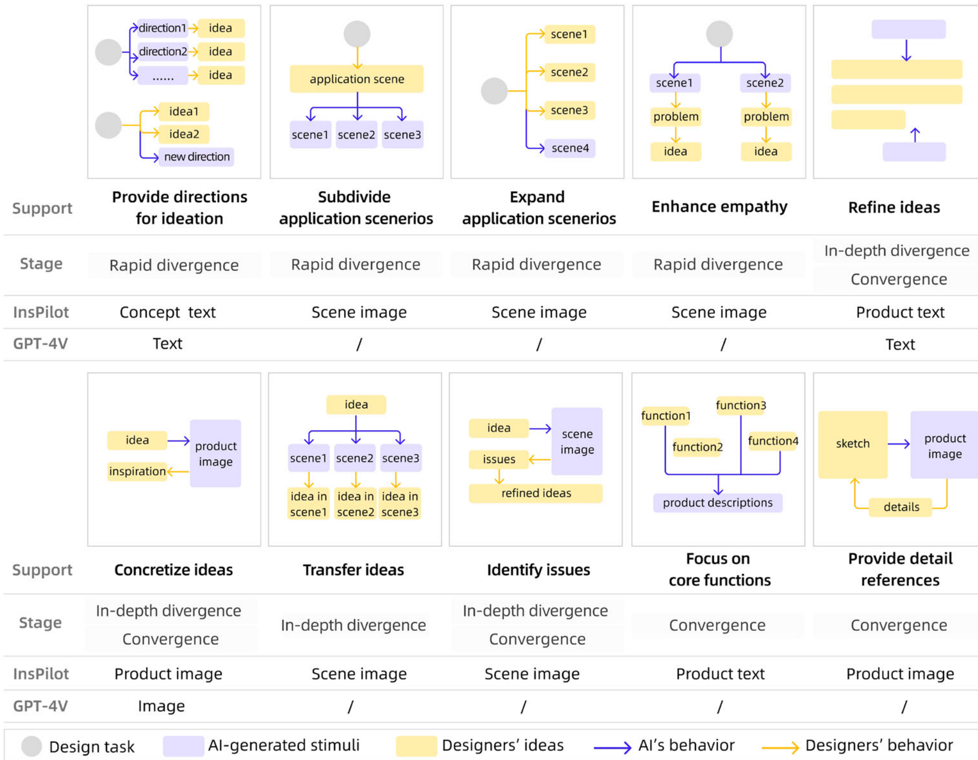

Figure 9. Creativity support provided by creative stimuli.

talked about general themes like disease prevention and environmental care'. For the visual stimuli generated by GPT-4V, 'the detailed product images occasionally led to my excessive focus on functional details' (P10).

In-depth Divergence Stage. At this stage, InsPilot's product texts and GPT-4V's texts helped participants refine ideas and supplement functional details. P12 mentioned that 'GPT added functions including utilising special materials and detecting road conditions to my design of a walking stick'. P1 also acknowledged the utility of product texts in supplementing unconsidered features, such as a disinfecting and sterilising function to a bathroom cleaning product. In addition, three participants felt that the product images generated by GPT-4V helped them concretize ideas and further inspired creativity.

The scene images generated by InsPilot further enhanced designers' ability to transfer ideas to multiple scenes. For instance, P1 initially designed a smart robot that could sweep and mop at the first stage, but after viewing the scene image of a bathroom, she adapted the idea to the bathroom scenario, customising features like water absorption and high-temperature disinfection. Moreover, an essential aspect of in-depth divergence is the selective expansion of existing ideas, where scene images aid designers in identifying potential issues within their ideas, thereby discarding less promising ideas and refining the others. For example, upon seeing an intersection scene, P10 realised that the initial design of the smart cane might not provide adequate mobility assistance for elderly users in such a setting. Similarly, when viewing scenes of the park and supermarket, P7 recognised that the wheelchair lift system she designed was impractical and abandoned it, as most scenes frequented by the elderly already have well-established accessibility systems.

However, for InsPilot, one participant felt that InsPilot's scene images were not closely related to his ideas and did not provide enough help. Regarding GPT-4V, two participants believed that the generated texts 'contained lots of fragmented functions, and I still needed to think about how to integrate them' (P4). They also noted that 'the generated texts were somewhat repetitive compared to the previous stage' (P2). Additionally, three participants did not use GPT-4V to generate texts because they 'needed to rephrase the ideas in the previous stage for GPT, and this process was too cumbersome; maybe I could ideate more deeply and quickly on my own' (P3).

Convergence Stage At this stage, InsPilot's and GPT-4V's textual stimuli provided different forms of support. Four participants believed that InsPilot helped them focus on core functions, saying that 'the generated texts categorised and integrated the functions in my ideas and eliminated unnecessary ones' (P5). Additionally, P10 commended InsPilot for transforming preliminary sketches and keywords into comprehensive product function descriptions. On the other hand, four participants utilised GPT-4V's generated texts to refine ideas, with P8 describing the texts as useful for improving selected ideas.

Furthermore, both InsPilot and GPT-4V's product images provided designers with forward incrementation (Sternberg 2005), which served to concretize ideas and provide detailed references. P3 stated that InsPilot was able to 'visually articulate the idea in my mind', while P2 appreciated GPT-4V's assistance in forming plausible product appearances. These images also offered valuable references for colour, material, and structural aspects. In addition, the scene images in InsPilot still facilitated issue identification in designers' existing ideas, which contributed to better idea selection and refinement. As an example, P4 mentioned that when considering practicality, one of her evaluation criteria was whether the product could cover as many application scenarios as possible with reference to the scene images provided by InsPilot.

However, images generated by both systems did not always meet participants' expectations. For instance, 'it (InsPilot) did not fully understand my ideas and sometimes generated contextually inappropriate images' (P11). Besides, P7 and P9 mentioned that 'its (GPT) generation results did not align with my expected results, and I was not allowed to further modify them'. Furthermore, two believed that the support from scene images diminished as they had already envisioned the scenarios by themselves. Besides, three participants criticised GPT-4V's text generation for its tendency to 'diverge uncontrollably' (P1), hindering their focus on the functions they wanted to converge on.

7.3. Influence on behaviour

Figure 10 presents an example of the behavioural coding results (more results in Appendix 3), including the time distribution of participants in the three stages, the timestamps for viewing creative stimuli/conversing with GPT, and the time periods for conceptualising ideas.

As calculated in Table 3, on average, participants spent less time using InsPilot during the first two stages compared to GPT-4V, but more time in the Convergence Stage. Combining the quantity of ideas in divergence stages (see Section 6.3), participants produced more ideas in a shorter time with InsPilot. The higher efficiency of ideation was not only

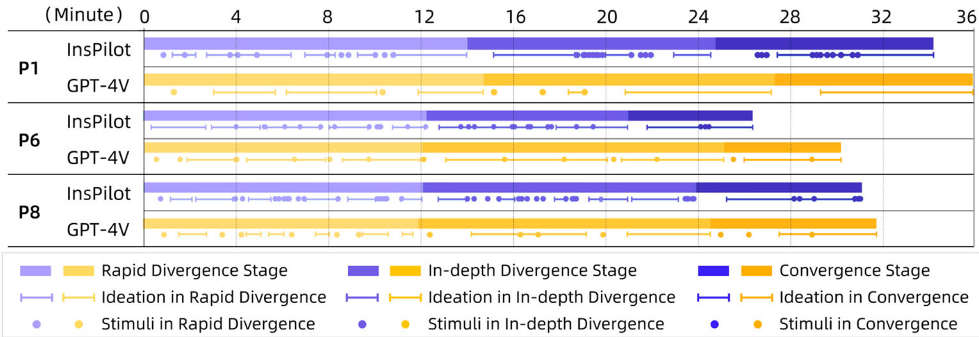

Figure 10. Example of the behaviour coding results.

Table 3. Descriptive statistics of the behaviour coding results.

		InsPilot		GPT-4V	
Criterion	Stage	M	SD	M	SD
Total time spent (s)	Rapid Divergence	790.3	116.0	837.8	112.0
	In-depth Divergence	605.2	82.0	650.0	115.6
	Convergence	455.8	92.2	407.9	100.4
Time spent before the first idea (s)	Rapid Divergence	75.0	48.7	172.3	74.4
	In-depth Divergence	54.2	24.8	109.0	82.2
	Convergence	120.8	55.0	115.1	78.8
Frequency of stimuli engagement (times)	Rapid Divergence	11.4	4.2	4.2	1.5
	In-depth Divergence	12.7	5.9	3.2	1.6
	Convergence	8.8	5.2	1.5	1.8

because InsPilot's intelligent generation strategy reduced the time participants spent communicating back and forth with the AI. More importantly, the stimuli generated by InsPilot were possibly more conducive to inspiring designers, which could be corroborated by more extensive support that InsPilot provided (Section 7.2).

In Rapid and In-depth Divergence, participants using InsPilot began sketching initial ideas earlier than those using GPT-4V, suggesting a greater fluency in ideation with InsPilot. Additionally, over half of the initial queries to GPT-4V in Rapid Divergence were for generating several design ideas, indicating a higher reliance on AI-generated ideas and potentially limiting personal creativity. Moreover, participants tended to get caught up in details during conversations with GPT-4V, gradually deviating from the divergent ideation process.

Regarding the frequency of stimuli engagement, on average, stimuli from InsPilot were viewed more frequently across all stages than stimuli from GPT-4V. InsPilot's proactive tracking of the design process and its capability to automatically generate stimuli based on participants' ideas reduced the cost of accessing stimuli, encouraging participants to actively view substantial stimuli. In contrast, participants needed to rephrase their design proposals to GPT-4V and ponder over wording, resulting in reduced enthusiasm for accessing stimuli.

Additionally, in the Rapid Divergence Stage, only six participants used GPT-4V to generate images; for In-depth Divergence, one participant abstained from using GPT-4V, and two participants did not use its image generation function; in Convergence, five participants did not generate new stimuli with GPT-4V but instead revisited stimuli from the previous two stages. This pattern indicates that many participants did not fully exploit GPT-4V for generating stimuli throughout the conceptual design process.

8. Discussion

8.1. Implications for generating differentiated multimodal stimuli

An innovative point of our research is the multiple modalities utilised in designers' input and in the generated stimuli. In InsPilot, the multimodal input transcends textual constraints and enables designers to sketch to convey ideas that can be challenging to articulate verbally (Ranscombe et al. 2024). For the generated stimuli, prior work also reveals that simultaneous presentation of textual and visual stimuli reduces text ambiguity and facilitates designers' exploration of images (Borgianni, Rotini, and Tomassini 2017; Shi et al. 2017). In this way, our strategy enables both manifestations of inspiration as mentioned in Ranscombe et al. (2024) including text-based generation and image-based creation, differentiating InsPilot from GPT-4V and grounding its potential for creativity support.

When applying the multimodal stimulus strategy, it is crucial to consider the personalised characteristics of each designer. Echoing Sternberg's insights (2005), designers exhibit individual variations in domain-specific design activities. According to our experiment, the effectiveness of stimuli also depends on the designer's proficiency to utilise information (Kavakli and Gero 2003), further modulated by factors like prior experience and the design method employed.

Firstly, designers with strong analogical thinking skills are commonly better at deriving ideas from abstract concepts (Yargın, Firth, and Crilly 2018). Conversely, others might struggle to extract usable information from brief texts. As Liang, Chang, and Liang (2019) demonstrated, abstract or surrealist visual stimuli provide remote analogues and are particularly effective in enhancing designers' intuition, exploration, and novelty generation. As a result, the differences between designers with different analogical thinking skills become apparent in the Rapid Divergence Stage, where scene images serve as distant yet related analogies. Similar to prior works that allow dynamic adjustments of semantic distance in creative stimuli (Althuizen and Wierenga 2014; X. T. Xu et al. 2021), future research could treat the abstraction level of creative stimuli as a continuous variable, introducing more details based on subtle individual differences.

Secondly, designers' familiarity with the design task typically affects their efficiency in utilising stimuli. We observed that when designers are well-acquainted with a task or have previously designed similar products, their behaviour may diverge into two extremes. Some may utilise stimuli that significantly differ from their prior experiences to broaden their thinking, while others were less inclined to adopt creative stimuli that differ from their own cognition, possibly due to the propensity to follow the path of least resistance in problem-solving (Ezzat et al. 2020). Such diversity in designers' expertise and experiences poses a challenge in providing universally applicable stimuli for all designers.

Apart from individual differences, the timing of stimulus introduction can significantly affect its effectiveness. Some stimuli may seem initially ineffective but can inspire serendipitous insights as the design process evolves (Mozaffari et al. 2022). For instance, P11 initially dismissed the textual stimulus 'enhancing ground friction to prevent slipping'. Later, while designing a multi-functional walking stick, this feature became crucial for ensuring safety. Conversely, the sunk cost effect implies that designers' reluctance to deviate from their initial ideas grows with invested time, potentially diminishing the positive impact of creative stimuli (Viswanathan and Linsey 2013). To sum up, future work could further investigate how creative stimuli affect individual designers differently at varying times.

8.2. Benefits and challenges of proactively generating creative stimuli

Throughout designers' interaction with InsPilot, GenAI was automatically prompted to produce multiple stimuli. The most apparent feature of this automation is its ability to prevent information loss between designers and GenAI through proactive prompt engineering (Y. Wang, Shen, and Lim 2023). In the field of conceptual design, a more distinctive advantage of proactive generation is that it mitigates homogeneity and enhances analogical potential of the stimuli (Ranscombe et al. 2024) and prevents designers from being distracted during the continuous process of idea generation and conceptualisation (Y. Han et al. 2024). Moreover, as the inference process of GenAI can break free from designers' current thinking pattern, the proactively generated stimuli can effectively prevent design fixation, which is also beneficial for designers with professional prompt engineering skills. Furthermore, the intelligent generation strategy also enables junior designers to conceive ideas or consider user scenarios that would otherwise be beyond their capabilities and experience and narrows their gap with experienced designers (Liang, Chang, and Liang 2019).

Despite the benefits, participants' feedback reveals that the inherent uncertainty (Li and Du 2017) of GenAI brings new challenges in generating creative stimuli that have received little attention in prior research. First, at Rapid Divergence Stage, designers initially form ideas and then seek to envision application scenarios through creative stimuli to identify problems and suggest improvements. However, AI's lack of real-world perception makes it challenging to reproduce authentic application scenarios. In the experiment, when P5 requested GPT for a description of how elderly people currently use wheelchairs to access public transportation, GPT failed to answer the question and instead said that 'the promotion and improvement of barrier-free travel are an ongoing process'. This underscores AI's difficulty in delivering specific, real-world contextual information, necessitating fine-tuning with more authentic data from the conceptual design field.

Prior to the In-depth Divergence and Convergence stages, designers rely on AI to aid in idea screening and evaluation by generating relevant examples, as they often compare new ideas against existing ones (Benedek et al. 2023). However, AI struggles with the ambiguous and fluctuating evaluation preferences of designers, lacking intuitive perceptions of criteria like novelty and practicality. As designers' criteria vary across stages and tasks, AI must adeptly adjust to these shifting preferences and dynamically modify its generation strategies.

After choosing ideas for further development, a significant challenge for AI here is accurately understanding the context of ideas and avoiding out-of-context outputs. Potential solutions include engaging in collaborative efforts with developers to fine-tune the GenAI model through abundant conceptual design data. Furthermore, the human-in-the-loop approach (P. Wang et al. 2019) and human-centred design method (Shneiderman 2020) could be beneficial. These approaches incorporate designers' feedback and opinions into the interaction pipeline, further aligning the generated stimuli to user expectations and needs.

For the Convergence Stage, AI-generated product images offer visual references for aspects like appearance and structure. However, participants found GenAI struggles in representing complex products combining multiple functionalities or with entirely new structures. Apart from that, the style of generated product images occasionally does not align with the target users or scenarios, which may arise from the current limitations of AI in fully comprehending the product styles and ambiance.

8.3. Limitations and future work

While our findings demonstrate InsPilot's potential, we acknowledge certain limitations stemming from both our experimental methodology and the inherent characteristics of the proposed strategy and current GenAI capabilities. We also identify avenues for future refinement of InsPilot.

Limitations of the Strategy and System. In terms of the proposed strategy and system, the provided stimuli aim to support problem identification and resolution across design stages, but not all design methodologies adhere to such a workflow. For example, in a technology-driven design process (B. Kim, Joines, and Feng 2023), the starting point often revolves around exploring potential needs based on technological capabilities rather than discovering user pain points from application scenarios.

Another shortcoming of InsPilot is its limitation to only textual and visual stimuli, and a few participants expressed their desire for stimuli of other modalities such as 3D models or videos (Urban Davis et al. 2021). However, since the current GenAI models have relatively limited speed and quality of generating stimuli in such modalities, they struggle to meet designers' substantial demand for high-quality stimuli in conceptual design. We will continuously monitor GenAI's advancement and incorporate more stimuli modalities when the technology matures.

Methodological Limitations. Methodologically, this paper evaluates InsPilot through a 32-minute user experiment, primarily focussing on the core rapid ideation and iterative nature of conceptual design, without longer experiment duration or assessment in real-world design contexts considering research practicality. In actual design activities, a designer's creativity can be influenced by myriad factors, including task familiarity, external environment conditions, and peer collaboration. Meanwhile, since designers mostly adopt sketches and texts to express ideas in the experiment, the validated effectiveness and the qualitative findings might not be suitable for tasks like physical prototyping.

Despite these constraints, we believe that the comparative user experiment meticulously replicates the quintessential stages of conceptual design, lending credibility to our findings. Future investigations will seek to evaluate InsPilot with more diverse tasks and in settings resembling authentic design situations, such as design workshops and field studies. Additionally, it is worth exploring in-depth that when using InsPilot, how designers with diverse backgrounds (e.g. industrial design, mechanical design, intelligent design) demonstrate differences in their creativity and utilisation of AI-generated creative stimuli.

Future Work and Design Implications. To further improve the generality and transferability of this pipeline, future endeavours could explore a modular framework, segmenting this pipeline into discrete and combinable units, so that the system could generate stimulus types tailored to empower other design processes and domains (e.g. emotional design, sustainable design). Furthermore, future work can regard the proposed pipeline as a minimal unit in the design process, repeatedly utilising it in real-world scenarios for adaptation to more sophisticated iterative processes.

Based on participants' feedback, we also found several underexplored but meaningful features for stimulus generation systems: (1) reusability of stimuli: several participants were limited by their relatively poor expressive abilities compared to AI when attempting to incorporate the product images into their own designs. Future research could adapt the interaction designs in some AI co-creation systems (Oh et al. 2018) to allow combinations of AI-generated content and users' sketches, thereby enhancing the reusability of stimuli and improving designers' expressiveness. (2) interpretability of stimuli: participants occasionally found it required effort to identify the connection between AI-generated stimuli and their own ideas, or they struggled to rapidly extract inspiring information from the stimuli. Future work could incorporate AI interpretability strategies (Kaur et al. 2022), providing more explanatory or guiding texts.

9. Conclusion

With the question of how to intelligently generate multimodal creative stimuli throughout the conceptual design process, this paper proposed the intelligent generation strategy consisting of prompt structures for generating various stimuli and an interaction pipeline to instruct the automatic generation process. Subsequently, we designed and developed InsPilot, an intelligent system capable of providing multimodal stimuli based on the design task, stage, design ideas, and the designer's input status. In the following comparative study, InsPilot was compared with GPT-4V to reveal quantitative and qualitative insights. In summary, this paper has made three significant contributions:

- The proposed interaction pipeline offers a practical and replicable approach for the intelligent generation of multimodal creative stimuli, extending prior work that focussed on certain stages or unimodal stimuli comprehensively.
- The developed system InsPilot supports sketch-based iterative prototyping. The generated stimuli adapt based on the design stage to enhance the flexibility of the prototyping process. In Rapid Divergence, InsPilot generates stimuli tailored to the current design task, helping designers create multiple parallel ideas without refining existing designs. It also facilitates the reflection, iteration, and development of original ideas during In-depth Divergence and helps designers filter ideas and iterate them towards the final solution in Convergence. The design of InsPilot has been proven more effective in fostering divergence and boosting designers' confidence compared to other AI-based sequential prototyping tools.
- The experiment findings delineate the impacts and implications associated with the intelligent generation of multimodal creative stimuli. Specifically, the creativity support provided by AI-generated stimuli including subdividing application scenarios and enhancing empathy in Rapid Divergence; refining, concretising, and transferring ideas in In-depth Divergence; identifying issues and focussing on core functions in Convergence, etc. Furthermore, our work reveals the two distinct systems' (InsPilot and GPT-4V) impact on designers' behaviour and their frequency of stimuli engagement.

Notes

1. https://www.pinterest.com/
2. https://chat.openai.com/
3. https://www.midjourney.com
4. https://stability.ai/
5. https://openai.com/dall-e-3
6. https://ant.design/

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Key Research and Development Program of China [2022YFB3303301].

References

Althuizen, N., and B. Wierenga. 2014. "Supporting Creative Problem Solving with a Case-Based Reasoning System." *Journal of Management Information Systems* 31 (1): 309–340. https://doi.org/10.2753/MIS0742-1222310112.

Andreasen, M. M., C. T. Hansen, and P. Cash. 2015. *Conceptual Design*. Cham, Switzerland: Springer.

Bangor, A., P. T. Kortum, and J. T. Miller. 2008. "An Empirical Evaluation of the System Usability Scale." *International Journal of Human–Computer Interaction* 24 (6): 574–594. https://doi.org/10.1080/10447310802205776.

Benami, O., and Y. Jin. 2008. "Creative Stimulation in Conceptual Design." In *American Society of Mechanical Engineers Digital Collection*, 251–263. Montreal, Quebec, Canada. https://doi.org/10.1115/DETC2002/DTM-34023.

Benedek, M., R. E. Beaty, D. L. Schacter, and Y. N. Kenett. 2023. "The Role of Memory in Creative Ideation." *Nature Reviews Psychology* 2:246–257. https://doi.org/10.1038/s44159-023-00158-z. number: 4.

Benton, L., G. Varotsis, and A. Vasalou. 2019. "Leading by Example: Exploring the Influence of Design Examples on Children's Creative Ideation." *International Journal of Human-Computer Studies*122:174–183. https://doi.org/10.1016/j.ijhcs.2018.09.007.

Borgianni, Y., F. Rotini, and M. Tomassini. 2017. "Fostering Ideation in the Very Early Design Phases: How Textual, Pictorial and Combined Stimuli Affect Creativity." In *DS 87-8 Proceedings of the 21st International Conference on Engineering Design (ICED 17) Vol 8: Human Behaviour in Design, Vancouver, Canada, 21–25.08.2017*, 139–148. Vancouver, Canada.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell. 2020. "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33:1877–1901. https://doi.org/10.48550/arXiv.2005.14165.

Chakrabarti, A., P. Sarkar, B. Leelavathamma, and B. S. Nataraju. 2005. "A Functional Representation for Aiding Biomimetic and Artificial Inspiration of New Ideas." *AI EDAM* 19:113–132. https://doi.org/10.1017/S0890060405050109.

Chan, J., K. Fu, C. Schunn, J. Cagan, K. Wood, and K. Kotovsky. 2011. "On the Benefits and Pitfalls of Analogies for Innovative Design: Ideation Performance Based on Analogical Distance, Commonness, and Modality of Examples." *Journal of Mechanical Design* 133: 081004. https://doi.org/10.1115/1.4004396.

Chen, L., Z. Cai, W. Cheang, Q. Long, L. Sun, P. Childs, and H. Zuo. 2025. "Asknaturegpt: An Llm-Driven Concept Generation Method Based on Bio-Inspired Design Knowledge." *Journal of Engineering Design*1–35.

Chen, L., W. Cheang, Z. Jiang, Y. Xu, Z. Cai, L. Sun, P. Childs, J. Han, P. Hansen, and H. Zuo. 2025. "I-Card: A Generative AI-Supported Intelligent Design Method Card Deck." In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–22. Yokohama, Japan.

Chen, P., Z. Li, Y. Wu, H. Zhang, J. Zhou, and L. Sun. 2024. "Modality and Fidelity: Understanding How Creative Stimulus Combinations Impact Design Outcomes and Process across Different Conceptual Design Phases." *Thinking Skills and Creativity* 54:101657. https://doi.org/10.1016/j.tsc.2024.101657.

Chen, P., Y. Wu, Z. Li, H. Zhang, M. Zhou, J. Yao, W. You, and L. Sun. 2025c. "Gpsdesign: Integrating Generative AI with Problem-Solution Co-Evolution Network to Support Product Conceptual Design." *International Journal of Human–Computer Interaction* 1–21.

Chen, P., J. Yao, Z. Cheng, Y. Cai, J. Li, W. You, and L. Sun. 2025d. "Coexploreds: Framing and Advancing Collaborative Design Space Exploration between Human and AI." In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–20. Yokohama, Japan.

Chen, Q., T. Zhang, M. Nie, Z. Wang, S. Xu, W. Shi, and Z. Cao. 2023. "Fashion-GPT: Integrating LLMs with Fashion Retrieval System." In *Proceedings of the 1st Workshop on Large Generative Models Meet Multimodal Applications*, 69–78. Ottawa, Canada. https://doi.org/10.1145/3607827.3616844.

Cherry, E., and C. Latulipe. 2014. "Quantifying the Creativity Support of Digital Tools through the Creativity Support Index." *ACM Transactions on Computer-Human Interaction* 21:1–25. https://doi.org/10.1145/2617588.

Chung, J. J. Y., and E. Adar. 2023. "PromptPaint: Steering Text-to-Image Generation through Paint Medium-Like Interactions." In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–17. San Francisco, CA, USA: Association for Computing Machinery. https://doi.org/10.1145/3586183.3606777.

Darzentas, D., R. Velt, R. Wetzel, P. J. Craigon, H. G. Wagner, L. D. Urquhart, and S. Benford. 2019. "Card Mapper: Enabling Data-Driven Reflections on Ideation Cards." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15. Glasgow, Scotland: Association for Computing Machinery. https://doi.org/10.1145/3290605.3300801.

Design Council. 2005. "The Double Diamond Process Model." https://www.designcouncil.org.uk/our-resources/the-double-diamond/.

Di Fede, G., D. Rocchesso, S. P. Dow, and S. Andolina. 2022. "The Idea Machine: LLM-based Expansion, Rewriting, Combination, and Suggestion of Ideas." In *Proceedings of the 14th Conference on Creativity and Cognition*, 623–627. Venice, Italy: Association for Computing Machinery. https://doi.org/10.1145/3527927.3535197.

Ezzat, H., M. Agogué, P. Le Masson, B. Weil, and M. Cassotti. 2020. "Specificity and Abstraction of Examples: Opposite Effects on Fixation for Creative Ideation." *The Journal of Creative Behavior* 54 (1): 115–122. https://doi.org/10.1002/jocb.349.

Fang, C., Y. Zhu, L. Fang, Y. Long, H. Lin, Y. Cong, and S. J. Wang. 2025. "Generative AI-Enhanced Human-AI Collaborative Conceptual Design: A Systematic Literature Review." *Design Studies* 97:101300. https://doi.org/10.1016/j.destud.2025.101300.

Feng, Y., X. Wang, K. K. Wong, S. Wang, Y. Lu, M. Zhu, B. Wang, and W. Chen. 2023. "Promptmagician: Interactive Prompt Engineering for Text-to-Image Creation." *IEEE Transactions on Visualization and Computer Graphics* 30 (1): 295–305. https://doi.org/10.1109/TVCG.2023.3327168.

Fogarty, J., D. Tan, A. Kapoor, and S. Winder. 2008. "CueFlik: Interactive Concept Learning in Image Search." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA*, 29–38. Florence, Italy: Association for Computing Machinery. https://doi.org/10.1145/1357054.1357061.

Goldschmidt, G., and A. L. Sever. 2011. "Inspiring Design Ideas with Texts." *Design Studies* 32 (2): 139–155. https://doi.org/10.1016/j.destud.2010.09.006.

Golembewski, M., and M. Selby. 2010. "Ideation Decks: A Card-Based Design Ideation Tool." In *Proceedings of the 8th ACM Conference on Designing Interactive Systems, New York, NY, USA*, 89–92. Aarhus, Denmark: Association for Computing Machinery. https://doi.org/10.1145/1858171.1858189.

Gomes, M. G., A. Ogliari, R. B. Fernandes, and K. O. Marques. 2022. "Evaluation of Physical Models as Creative Stimuli in Conceptual Design of Products." *Design Studies* 81:101119. https://doi.org/10.1016/j.destud.2022.101119.

Gonalves, M., C. Cardoso, and P. Badke-Schaub. 2012. "Find Your Inspiration: Exploring Different Levels of Abstraction in Textual Stimuli." In *DS 73-1 Proceedings of the 2nd International Conference on Design Creativity Volume 1*. Glasgow, UK.

Han, Y., Z. Qiu, J. Cheng, and R. Lc. 2024. "When Teams Embrace AI: Human Collaboration Strategies in Generative Prompting in a Creative Design Task." In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–14. Honolulu, Hawaii, USA.

Han, J., F. Shi, L. Chen, and P. R. N. Childs. 2018. "A Computational Tool for Creative Idea Generation Based on Analogical Reasoning and Ontology." *AI EDAM* 32:462–477. https://doi.org/10.1017/S0890060418000082.

Hatcher, G., W. Ion, R. Maclachlan, M. Marlow, B. Simpson, N. Wilson, and A. Wodehouse. 2018. "Using Linkography to Compare Creative Methods for Group Ideation." *Design Studies* 58:127–152. https://doi.org/10.1016/j.destud.2018.05.002.

Herring, S. R., C. C. Chang, J. Krantzler, and B. P. Bailey. 2009. "Getting Inspired! Understanding How and why Examples Are Used in Creative Design Practice." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 87–96. Boston, Massachusetts, USA: Association for Computing Machinery. https://doi.org/10.1145/1518701.1518717.

Hou, Y., X. Zeng, Y. Wang, M. Yang, X. Chen, and W. Zeng. 2025. "Gencolor: Generative Color-Concept Association in Visual Design." In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–19. Yokohama, Japan.

Huang, L., F. Gino, and A. D. Galinsky. 2015. "The Highest Form of Intelligence: Sarcasm Increases Creativity for Both Expressers and Recipients." *Organizational Behavior and Human Decision Processes*131:162–177. https://doi.org/10.1016/j.obhdp.2015.07.001.

Jin, X., M. Evans, H. Dong, and A. Yao. 2021. "Design Heuristics for Artificial Intelligence: Inspirational Design Stimuli for Supporting UX Designers in Generating AI-Powered Ideas." In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–8. https://doi.org/10.1145/3411763.3451727.

Jin, J., M. Yang, H. Hu, X. Guo, J. Luo, and Y. Liu. 2025. "Empowering Design Innovation Using AI-Generated Content." *Journal of Engineering Design* 36 (1): 1–18.

Kang, Y., Z. Sun, S. Wang, Z. Huang, Z. Wu, and X. Ma. 2021. "MetaMap: Supporting Visual Metaphor Ideation through Multi-dimensional Example-Based Exploration." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15. Association for Computing Machinery. https://doi.org/10.1145/3411764.3445325.

Karimi, P., J. Rezwana, S. Siddiqui, M. L. Maher, and N. Dehbozorgi. 2020. "Creative Sketching Partner: An Analysis of Human-AI Co-Creativity." In *Proceedings of the 25th International Conference on Intelligent User Interfaces, New York, NY, USA*, 221–230. Cagliari, Italy: Association for Computing Machinery. https://doi.org/10.1145/3377325.3377522.

Kaur, H., E. Adar, E. Gilbert, and C. Lampe. 2022. "Sensible AI: Re-Imagining Interpretability and Explainability Using Sensemaking Theory." In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 702–714. Seoul, South Korea: https://doi.org/10.1145/3531146.3533135.

Kavakli, M., and J. S. Gero. 2003. "Strategic Knowledge Differences between an Expert and a Novice Designer." In *Human Behaviour in Design: Individuals, Teams, Tools*, 42–52. Springer. https://doi.org/10.1007/978-3-662-07811-2_5.

Kim, B., S. Joines, and J. Feng. 2023. "Technology-Driven Design Process: Teaching and Mentoring Technology-Driven Design Process in Industrial Design Education." *International Journal of Technology and Design Education* 33 (2): 521–555. https://doi.org/10.1007/s10798-022-09739-0.

Kim, J., and M. L. Maher. 2023. "The Effect of AI-based Inspiration on Human Design Ideation." *International Journal of Design Creativity and Innovation* 11 (2): 81–98. https://doi.org/10.1080/21650349.2023.2167124.

Kim, J., M. Maher, and S. Siddiqui. 2021. "Studying the Impact of AI-Based Inspiration on Human Ideation in a Co-Creative Design System." In *IUI Workshops*.

Koronis, G., A. Silva, J. K. K. Siang, and C. Yogiaman. 2022. "A Study on the Link between Design Brief Structure and Stimulus Fidelity to Optimize Novelty and Usefulness." *AI EDAM* 36:e5. https://doi.org/10.1017/S0890060421000378.

Kuys, B., C. Ranscombe, and W. Zhang. 2023. "Visualising Product Concepts to Engage Manufacturers with Little or No Industrial Design Capability." *Design Studies* 87:101188. https://doi.org/10.1016/j.destud.2023.101188.

Lee, T., M. Yasunaga, C. Meng, Y. Mai, J. S. Park, A. Gupta, Y. Zhang, D. Narayanan, H. Teufel, M. Bellagente, et al. 2024. "Holistic Evaluation of Text-to-Image Models." In *Advances in Neural Information Processing Systems 36*. New Orleans, Louisiana, USA. https://doi.org/10.48550/arXiv.2311.04287.

Li, D., and Y. Du. 2017. *Artificial Intelligence with Uncertainty*. CRC Press. https://doi.org/10.1201/9781315366951.

Liang, C. T., Y. Chang, and C. Liang. 2019. "How Can Pictorial Representations Stimulate the Imaginative Capacity of Experienced Multimedia Designers?" *International Journal of Design Creativity and Innovation* 7 (3): 179–192. https://doi.org/10.1080/21650349.2018.1465477.

Lin, D. C. E., H. B. Kang, N. Martelaro, A. Kittur, Y. Y. Chen, and M. K. Hong. 2025. "Inkspire: Supporting Design Exploration with Generative AI through Analogical Sketching." In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–18. Yokohama, Japan.

Lindley, S., X. Cao, J. Helmes, R. Morris, and S. Meek. 2013. "Towards a Tool for Design Ideation: Insights from Use of SketchStorm." In *Proceedings of the 27th BCS Conference on Human Computer Interaction*. London, UK.

Liu, W., A. Dempo, T. Kimura, T. Kawashima, and K. Shinohara. 2022. "Effects of the Presence of a Cell Phone and Exposure to Natural Environments on Remote Associates Task Performance." *Scientific Reports* 12:9507. https://doi.org/10.1038/s41598-022-13634-y.

Liu, P., W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. 2023. "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing." *ACM Computing Surveys*55:1–35. https://doi.org/10.1145/3560815.

Lomas, J. D., M. Karac, and M. Gielen. 2021. "Design Space Cards: Using a Card Deck to Navigate the Design Space of Interactive Play." *Proceedings of the ACM on Human-Computer Interaction* 5:1–21. https://doi.org/10.1145/3474654.

Mora, S., F. Gianni, and M. Divitini. 2017. "Tiles: A Card-Based Ideation Toolkit for the Internet of Things." In *Proceedings of the 2017 Conference on Designing Interactive Systems, New York, NY, USA*, 587–598. Edinburgh, UK: Association for Computing Machinery. https://doi.org/10.1145/3064663.3064699.

Mozaffari, M. A., X. Zhang, J. Cheng, and J. L. Guo. 2022. "GANSpiration: Balancing Targeted and Serendipitous Inspiration in User Interface Design with Style-Based Generative Adversarial Network." In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–15. New Orleans, Louisiana: Association for Computing Machinery. https://doi.org/10.1145/3491102.3517511.

Nijstad, B. A., W. Stroebe, and H. F. M. Lodewijkx. 2002. "Cognitive Stimulation and Interference in Groups: Exposure Effects in an Idea Generation Task." *Journal of Experimental Social Psychology* 38 (6): 535–544. https://doi.org/10.1016/S0022-1031(02)00500-0.

Oh, C., J. Song, J. Choi, S. Kim, S. Lee, and B. Suh. 2018. "I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-creation with Artificial Intelligence." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. Montréal, Canada. https://doi.org/10.1145/3173574.3174223.

Otani, M., R. Togashi, Y. Sawai, R. Ishigami, Y. Nakashima, E. Rahtu, J. Heikkilä, and S. Satoh. 2023. "Toward Verifiable and Reproducible Human Evaluation for Text-to-Image Generation." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14277–14286. Seattle, Washington. https://doi.org/10.1109/CVPR52729.2023.01372.

Paay, J., J. Kjeldskov, M. Bønnerup, and T. Rasenthiran. 2023. "Sketching and Context: Exploring Creativity in Idea Generation Groups." *Design Studies* 84:101159. https://doi.org/10.1016/j.destud.2022.101159.

Pei, E., I. Campbell, and M. Evans. 2011. "A Taxonomic Classification of Visual Design Representations Used by Industrial Designers and Engineering Designers." *The Design Journal* 14 (1): 64–91. https://doi.org/10.2752/175630610X12877385838803.

Peng, X., J. Koch, and W. E. Mackay. 2025. "Fusain: Composing Generative AI Visual Prompts Using Pen-Based Interaction." In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–20. Yokohama, Japan.

Podell, D., Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. 2023. "SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis." *arXiv preprint* arXiv:2307.01952. https://doi.org/10.48550/arXiv.2307.01952.

Ranscombe, C., L. Tan, M. Goudswaard, and C. Snider. 2024. "Inspiration or Indication? Evaluating the Qualities of Design Inspiration Boards Created Using Text to Image Generative Ai." *Proceedings of the Design Society* 4:2207–2216. https://doi.org/10.1017/pds.2024.223.

Rombach, R., A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. 2022. "High-Resolution Image Synthesis with Latent Diffusion Models." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695. New Orleans, Louisiana. https://doi.org/10.1109/CVPR52688.2022.01042.

Ruiz-Pastor, L., Y. Borgianni, and V. Chulvi. 2024. "Sustainability-Oriented Stimuli and Their Effects on Circularity and Novelty in Conceptual Design." *International Journal of Design Creativity and Innovation* 12 (2): 81–100. https://doi.org/10.1080/21650349.2024.2316915.

Sarkar, P., and A. Chakrabarti. 2011. "Assessing Design Creativity." *Design Studies* 32 (4): 348–383. https://doi.org/10.1016/j.destud.2011.01.002.

Sassenberg, K., G. B. Moskowitz, A. Fetterman, and T. Kessler. 2017. "Priming Creativity as a Strategy to Increase Creative Performance by Facilitating the Activation and Use of Remote Associations." *Journal of Experimental Social Psychology* 68:128–138. https://doi.org/10.1016/j.jesp.2016.06.010.

Setchi, R., and C. Bouchard. 2010. "In Search of Design Inspiration: A Semantic-Based Approach." *Journal of Computing and Information Science in Engineering* 10. https://doi.org/10.1115/1.3482061.

Shi, Y., Y. Wang, Y. Qi, J. Chen, X. Xu, and K. L. Ma. 2017. "IdeaWall: Improving Creative Collaboration through Combinatorial Visual Stimuli." In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, New York, NY, USA*, 594–603. Portland, Oregon, USA. https://doi.org/10.1145/2998181.2998208.

Shneiderman, B. 2020. "Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy." *International Journal of Human–Computer Interaction* 36 (6): 495–504. https://doi.org/10.1080/10447318.2020.1741118.

Siangliulue, P., J. Chan, K. Z. Gajos, and S. P. Dow. 2015. "Providing Timely Examples Improves the Quantity and Quality of Generated Ideas." In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, 83–92. Seoul, South Korea: Association for Computing Machinery. https://doi.org/10.1145/2757226.2757230.

Sternberg, R. J. 2003. *Wisdom, Intelligence, and Creativity Synthesized*. Cambridge University Press.

Sternberg, R. J. 2005. "Creativity or Creativities?" *International Journal of Human-Computer Studies*63 (4–5): 370–382. https://doi.org/10.1016/j.ijhcs.2005.04.003.

Sun, Z., Z. Zhang, Y. Zhang, M. Lu, D. Lischinski, D. Cohen-Or, and H. Huang. 2025. "Creative Blends of Visual Concepts." In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–17. Yokohama, Japan.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M. A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, and F. Azhar, et al. 2023. "LLaMA: Open and Efficient Foundation Language Models." arXiv:2302.13971.

Urban Davis, J., F. Anderson, M. Stroetzel, T. Grossman, and G. Fitzmaurice. 2021. "Designing Co-Creative AI for Virtual Environments." In *Proceedings of the 13th Conference on Creativity and Cognition, New York, NY, USA*. Association for Computing Machinery. https://doi.org/10.1145/3450741.3465260.

Vandevenne, D., T. Pieters, C. Vanneste, and J. R. Duflou. 2016. "Case-Dependence of Level-Based Idea Space Variety for Systematic Biologically-Inspired Design." *Procedia CIRP* 39:173–178. https://doi.org/10.1016/j.procir.2016.01.184.

Van Noorden, R., and J. M. Perkel. 2023. "AI and Science: What 1,600 Researchers Think." *Nature* 621 (7980): 672–675. https://doi.org/10.1038/d41586-023-02980-0.

Vasconcelos, L. A., and N. Crilly. 2016. "Inspiration and Fixation: Questions, Methods, Findings, and Challenges." *Design Studies* 42:1–32. https://doi.org/10.1016/j.destud.2015.11.001.

Vasconcelos, L., J. Zahn, D. Trevisan, and J. Viterbo. 2024. "Engagement by Design Cards: A Tool to Involve Designers and Non-Experts in the Design of Crowdsourcing Initiatives." *International Journal of Human-Computer Studies* 182:103166. https://doi.org/10.1016/j.ijhcs.2023.103166.

Vattam, S., B. Wiltgen, M. Helms, A. K. Goel, and J. Yen. 2011. "DANE: Fostering Creativity in and through Biologically Inspired Design." In *Design Creativity 2010*, 115–122. Springer. https://doi.org/10.1007/978-0-85729-224-7_16.

Verheijden, M. P., and M. Funk. 2023. "Collaborative Diffusion: Boosting Designerly Co-Creation with Generative AI." In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–8. Hamburg, Germany: Association for Computing Machinery. https://doi.org/10.1145/3544549.3585680.

Viswanathan, V. K., and J. S. Linsey. 2013. "Role of Sunk Cost in Engineering Idea Generation: An Experimental Investigation." *Journal of Mechanical Design* 135. https://doi.org/10.1115/1.4025290.

Wang, P., D. Peng, L. Li, L. Chen, C. Wu, X. Wang, P. Childs, and Y. Guo. 2019. "Human-in-the-Loop Design with Machine Learning." In *Proceedings of the Design Society: International Conference on Engineering Design*, 2577–2586. Cambridge University Press. https://doi.org/10.1017/dsi.2019.264.

Wang, S., S. Petridis, T. Kwon, X. Ma, and L. B. Chilton. 2023. "PopBlends: Strategies for Conceptual Blending with Large Language Models." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–19. Hamburg, Germany: Association for Computing Machinery. https://doi.org/10.1145/3544548.3580948.

Wang, Y., S. Shen, and B. Y. Lim. 2023. "Reprompt: Automatic Prompt Editing to Refine AI-Generative Art towards Precise Expressions." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–29. Hamburg, Germany:

Wu, T., E. Jiang, A. Donsbach, J. Gray, A. Molina, M. Terry, and C. J. Cai. 2022. "PromptChainer: Chaining Large Language Model Prompts through Visual Programming." In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems, New York, NY, USA*, 1–10. New Orleans, Louisiana, USA: https://doi.org/10.1145/3491101.3519729.

Wu, D., Z. Yu, N. Ma, J. Jiang, Y. Wang, G. Zhou, H. Deng, and Y. Li. 2023. "Styleme: Towards Intelligent Fashion Generation with Designer Style." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–16. Hamburg, Germany. https://doi.org/10.1145/3544548.3581377.

Xu, J., C. J. Chao, and Z. Fu. 2020. "Research on Intelligent Design Tools to Stimulate Creative Thinking." In *Cross-Cultural Design. User Experience of Products, Services, and Intelligent Environments*, 661–672. https://doi.org/10.1007/978-3-030-49788-0_50.

Xu, J., X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong. 2024. "Imagereward: Learning and Evaluating Human Preferences for Text-to-Image Generation." In *Advances in Neural Information Processing Systems 36*. New Orleans, Louisiana. https://doi.org/10.48550/arXiv.2304.05977.

Xu, X. T., R. Xiong, B. Wang, D. Min, and S. P. Dow. 2021. "IdeateRelate: An Examples Gallery That Helps Creators Explore Ideas in Relation to Their Own." *Proceedings of the ACM on Human-Computer Interaction* 5:1–18. https://doi.org/10.1145/3479496.

Yamamoto, Y., and K. Nakakoji. 2005. "Interaction Design of Tools for Fostering Creativity in the Early Stages of Information Design." *International Journal of Human-Computer Studies* 63 (4–5): 513–535. https://doi.org/10.1016/j.ijhcs.2005.04.023.

Yao, J., P. Chen, Z. Li, Y. Cai, Y. Wu, W. You, and L. Sun. 2025. "Stepideator: Utilizing Mixed Representations to Support Step-by-Step Design with Generative Artificial Intelligence." *Journal of Mechanical Design* 147 (7): 071703. https://doi.org/10.1115/1.4067426.

Yargın, G. T., R. M. Firth, and N. Crilly. 2018. "User Requirements for Analogical Design Support Tools: Learning from Practitioners of Bio-Inspired Design." *Design Studies* 58:1–35. https://doi.org/10.17863/CAM.20186.

Yee, K. P., K. Swearingen, K. Li, and M. Hearst. 2003. "Faceted Metadata for Image Search and Browsing." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA*, 401–408. Ft. Lauderdale, Florida, USA: Association for Computing Machinery. https://doi.org/10.1145/642611.642681.

Yilmaz, S., S. R. Daly, C. M. Seifert, and R. Gonzalez. 2016. "Evidence-Based Design Heuristics for Idea Generation." *Design Studies* 46:95–124. https://doi.org/10.1016/j.destud.2016.05.001.

Yuan, H., M. Liu, K. Lu, C. Yang, and N. Hao. 2023. "The Effect of Example Abstraction on Creativity from the Perspectives of Example Modality and Generality." *Thinking Skills and Creativity* 47:101234. https://doi.org/10.1016/j.tsc.2023.101234.

Yun, G., K. Cho, Y. Jeong, and T. J. Nam. 2022. "Ideasquares: Utilizing Generative Text as a Source of Design Inspiration." In *DRS Biennial Conference Series*. https://doi.org/10.21606/drs.2022.484.
Zhang, H., P. Chen, X. Xie, Z. Jiang, Y. Wu, Z. Li, X. Chen, and L. Sun. 2025. "Fusionprotor: A Mixed-Prototype Tool for Component-Level Physical-to-Virtual 3D Transition and Simulation." In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–19. Yokohama, Japan.
Zhang, H., P. Chen, X. Xie, C. Lin, L. Liu, Z. Li, W. You, and L. Sun. 2024. "Protodreamer: A Mixed-Prototype Tool Combining Physical Model and Generative AI to Support Conceptual Design." In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, 1–18. Pittsburgh, Pennsylvania, USA.
Zhao, W. X., K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, et al. 2023. "A Survey of Large Language Models." https://doi.org/10.48550/arXiv.2303.18223.
Zhou, M., D. Zhang, W. You, Z. Yu, Y. Wu, C. Pan, H. Liu, T. Lao, and P. Chen. 2024. "Stylefactory: Towards Better Style Alignment in Image Creation through Style-Strength-based Control and Evaluation." In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, 1–15. Pittsburgh, Pennsylvania, USA.
Zhou, C., X. Zhang, and C. Yu. 2023. "How Does AI Promote Design Iteration? The Optimal Time to Integrate AI into the Design Process." *Journal of Engineering Design*1–28. https://doi.org/10.1080/09544828.2023.2290915.

Appendix 1. Prompt design

A.1 Evaluation criteria of the generated stimuli

Before designing the prompt for generating different stimuli, we developed certain criteria with reference to relevant work (J. Xu et al. 2024; Lee et al. 2024; Otani et al. 2023) to evaluate whether the generated stimuli meet our requirement. The criteria and their description are shown in Table A1.

Table A1. Evaluation criteria for prompt iteration.

Type	Criterion	Description
All Types	Instruction Compliance	Whether the generated stimuli follow the instruction on the output format.
	Content Relevance	Whether the generated stimuli relevant to the input task or design ideas.
	Output Stability	Whether consistent quality results can be produced when executed 10 times.
Concept Text	Definition Compliance	Whether the stimuli meet the definition of concept texts (i.e. including only abstract definitions without specific product categories or functions).
	Diversity	Whether the concepts of different creative stimuli are non-repetitive, diverse, and can guide designers to think in different directions.
Product Text	Definition Compliance	Whether the stimuli meet the definition of product texts (i.e. introducing specific product functions and including at least three product features).
	Diversity	Whether the product categories of different creative stimuli are non-repetitive, with significant differences in the features of different products
Scene Image	Definition Compliance	Whether the image depicts a scene without obvious human figures.
	Rationality	Whether the image depicts a realistic scene with proper lighting, proportions, and shapes.
Product Image	Definition Compliance	Whether the image depicts a product without obvious human figures, and the background is clean without specific scenes.
	Rationality	Whether the product in the image adheres to realistic lighting, perspective, and proportions.

A.2 Iteration process of prompts

In this part, we introduce more details about the process of prompt design. The prompt for idea understanding is presented below, and texts in *italics* are alternative words that will be replaced according to the design task and input texts.

> Design Task: The sketch in this image is related to '*designing a sports assistance product*'.
> Product Functions: Its functions include *timing and counting.*
> Task Brief: Based on the sketch and functions, please tell me what product it is?
> Task Requirement: The product name should be between 5–10 words.
> Output Requirement: Only the product name should be provided, without numbers, quotes, or any other text or symbols.
> Example: Output Example – Portable Foldable Treadmill

The prompts for generating stimuli of product texts are presented below.

> Task Brief: I will provide you with one design task enclosed in < > brackets, and please generate ten texts of product descriptions for me.
> Task Requirement: The following requirements should be met:
> Word Limit: (1) The output should be between 20–50 words.
> (2) It should include detailed functions and structure of the product.
> Quantitative Requirement: (3) It should contain at least 3 or more feature points with comprehensive content.
> Example: For example, if the design task is to create a mobility aid product, a possible description could be: 'A cordless jump rope with adjustable length for personalised use. The handle features weighted balls for enhanced fat-burning effects. Equipped with timing and counting functions to record workout data'.
> Output Requirement: Please only output the prompts' content, one per line, without numbering or any other text!
> Limiting Symbols: Design Task – < *Design a product to assist individuals with disabilities in mobility* >

Figure A11 and A12 illustrate the iteration process of prompts for generating concept texts and product texts separately.

SDXL utilises positive prompts to detail desired content and negative prompts to exclude unwanted elements. Moreover, positive prompts include fixed prompts restricting image styles and adjustable prompts describing elements within images. Table A2 shows the positive and negative prompts for scene images and product images. Texts in *italics* represent adjustable prompts.

Figure A13 compares the quality of generated scene images with no fixed prompts, only fixed positive prompts, and both fixed positive and negative prompts. When there are no fixed prompts or only positive prompts, the generated images contain figures and unidentified shapes. When both positive and negative prompts are used, the generated images are more realistic and meet our requirements for scene images.

The generated product images based on different prompt combinations are shown in Figure A14. Without fixed prompts, the generated images include figures and the background is cluttered, containing actual scenarios. With only fixed positive prompts, although the background of the generated images is cleaner, there are some illogical elements. For example, the screen is put at the back of the treadmill. Additionally, the image quality and lighting effects are not exquisite enough. When using both positive and negative prompts, the generated images meet the requirements for product image stimuli.

Referring to the generation steps of scene images, we divide the generation process of product images into two steps, as shown in Figure A15.

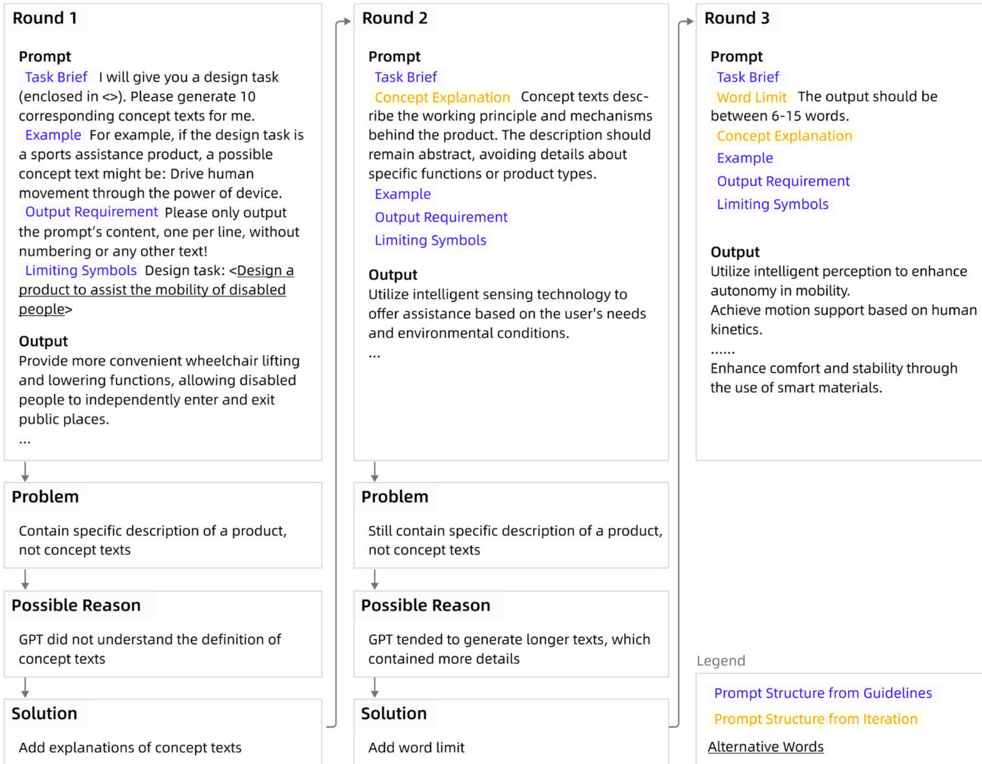

Figure A11. Iteration process of the prompts for generating concept texts.

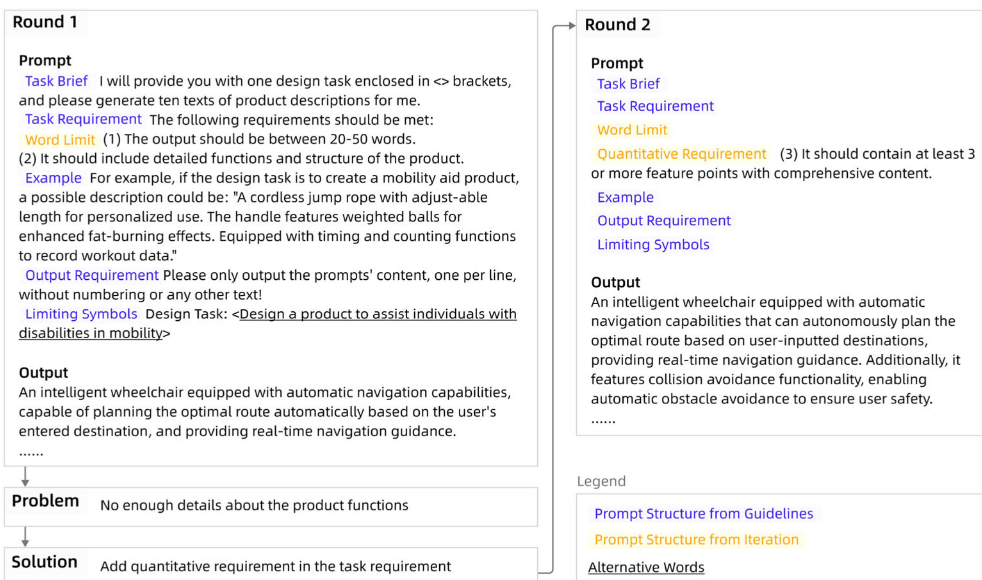

Figure A12. Iteration process of the prompts for generating product texts.

Table A2. SDXL prompts for generating images.

Image type	Prompt type	Content
Scene image	Positive prompts	(photorealistic:1.5), bestquality, ultradetailed, masterpiece, realistic, finely detailed, purism, minimalism, 4k, *home, gym, fitness, equipment, weights, treadmill, mirror, bright, indoor*
	Negative prompts	(worst quality:1.4), people, man, woman, flame, Cloud, (low quality:1.4), (normal quality:1.5), lowres, ((monochrome)), ((grayscale)), cropped, text, jpeg, artefacts, signature, watermark, username, sketch, cartoon, drawing, duplicate, anime, blurry, semi-realistic, out of frame, ugly, deformed, weird colours, EasyNegative, flame
Product image	Positive prompts	(white background:1.5), Actual product pictures, (Product Design:1.3), intelligent, industrial products, Creative, Industrial Products, sense of future, complete view, High Quality, ue 5, minimalistic futuristic design, emauromin style,finely detailed, 64k, blender, purism, minimalism, photorealistic, *treadmill, sleek design, metallic finish, touchscreen, adjustable lighting*
	Negative prompts	(worst quality:1.4), people, man, woman, flame, Cloud, (low quality:1.4), (normal quality:1.5), lowres, ((monochrome)), ((grayscale)), cropped, text, jpeg, artefacts, signature, watermark, username, sketch, cartoon, drawing, duplicate, anime, blurry, semi-realistic, out of frame, ugly, deformed, weird colours, Easy Negative, flame

Appendix 2. Design tasks in user experiment

In the user experiment in Section 6, Task 1 was to design a product to assist senior citizens in their mobility, and Task 2 was to design a smart household cleaning product. We developed these two tasks based on the same standards as the formative study and referenced prior work. The detailed design tasks provided to participants are shown below. We ensured that the task descriptions were similar in word count and structure, including a brief background introduction, target users' pain points, and design requirements.

> Task 1: Design a product to assist the senior citizens in their mobility
> Task Description: Senior citizens often face various mobility challenges, such as difficulty in walking, decreased sense of direction, and deterioration of vision and hearing. Senior citizens also wish to travel freely, but due to the above difficulties, they may feel worried or scared about going out. To solve these problems, your task is to design a product to assist senior citizens in their mobility.
> Task 2: Design a smart household cleaning product
> Task Description: In modern society, being busy has become the norm for many families. For many households, domestic cleaning is often time-consuming and tedious, especially in families with children or pets. Cleaning not only requires the investment of time but also continuous energy and physical labour. To free family members from burdensome cleaning tasks, your task is to design a smart household cleaning product.

Appendix 3. Behavior coding results

Figure A16 depicts the results of the behaviour coding of all participants.

To verify whether participants' ratings at different stages were influenced by their time allocation, we further conducted Kendall's correlation analysis, and the results are shown in Table A3.

According to Kendall's correlation analysis, an absolute value of Kendall's correlation coefficient less than 0.3 indicates no correlation between time and ratings. An absolute value between 0.3 and 0.7 with a p-value less than 0.05 indicates a significant weak correlation. An absolute value greater than 0.7 with a p-value less than 0.05 indicates a significant strong correlation. In our analysis results,

Figure A13. Generated scene images with different prompts.

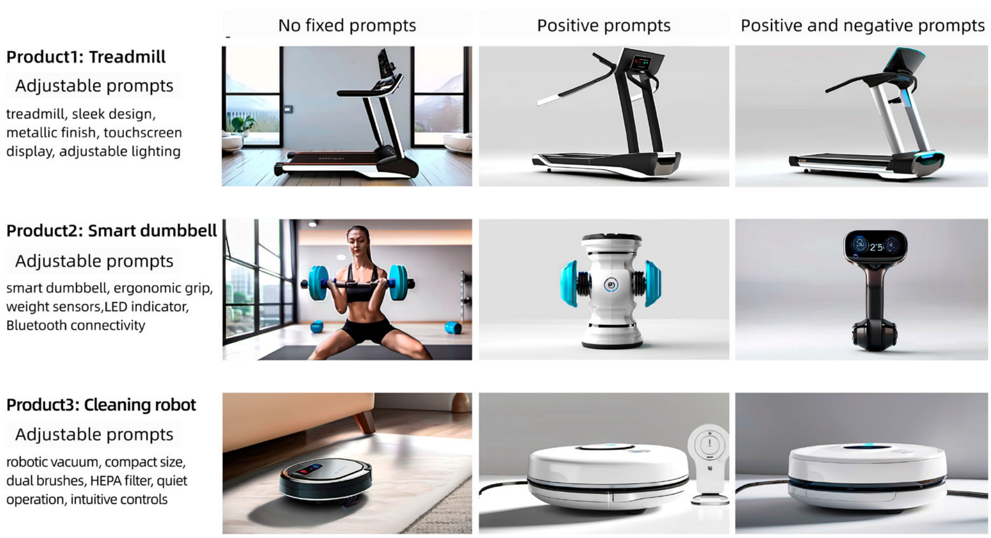

Figure A14. Generated product images with different prompts.

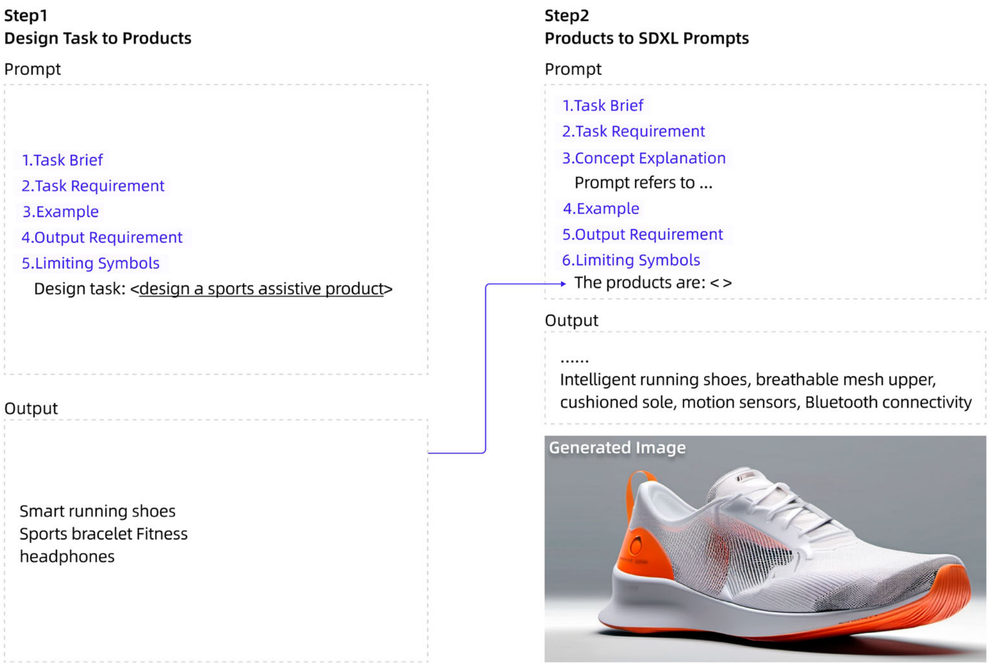

Figure A15. Process and prompts for generating product images.

Table A3. Results of the Kendall's correlation analysis between participants' ratings and their time allocation.

		Novelty		Practicality		Diversity	
	Stage	Kendall coefficient	*p*-value	Kendall coefficient	*p*-value	Kendall coefficient	*p*-value
InsPilot	Rapid Divergence	0.28	0.23	−0.23	0.33	−0.05	0.83
	In-depth Divergence	−0.24	0.33	−0.28	0.26	−0.31	0.19
	Convergence	−0.22	0.37	0.15	0.55	/	/
GPT-4V	Rapid Divergence	−0.12	0.61	−0.29	0.25	0.02	0.94
	In-depth Divergence	0.11	0.65	0.16	0.51	0.00	1.00
	Convergence	0.13	0.57	−0.09	0.70	/	/

Kendall's correlation coefficient was less than 0.3 for most dimensions. The correlation coefficient between the time spent in In-depth Divergence Stage of InsPilot and the diversity rating was −0.31, but with a *p*-value of 0.19. Therefore, there was no significant correlation between time and ratings, suggesting that participants' time allocation did not significantly influence their ratings.

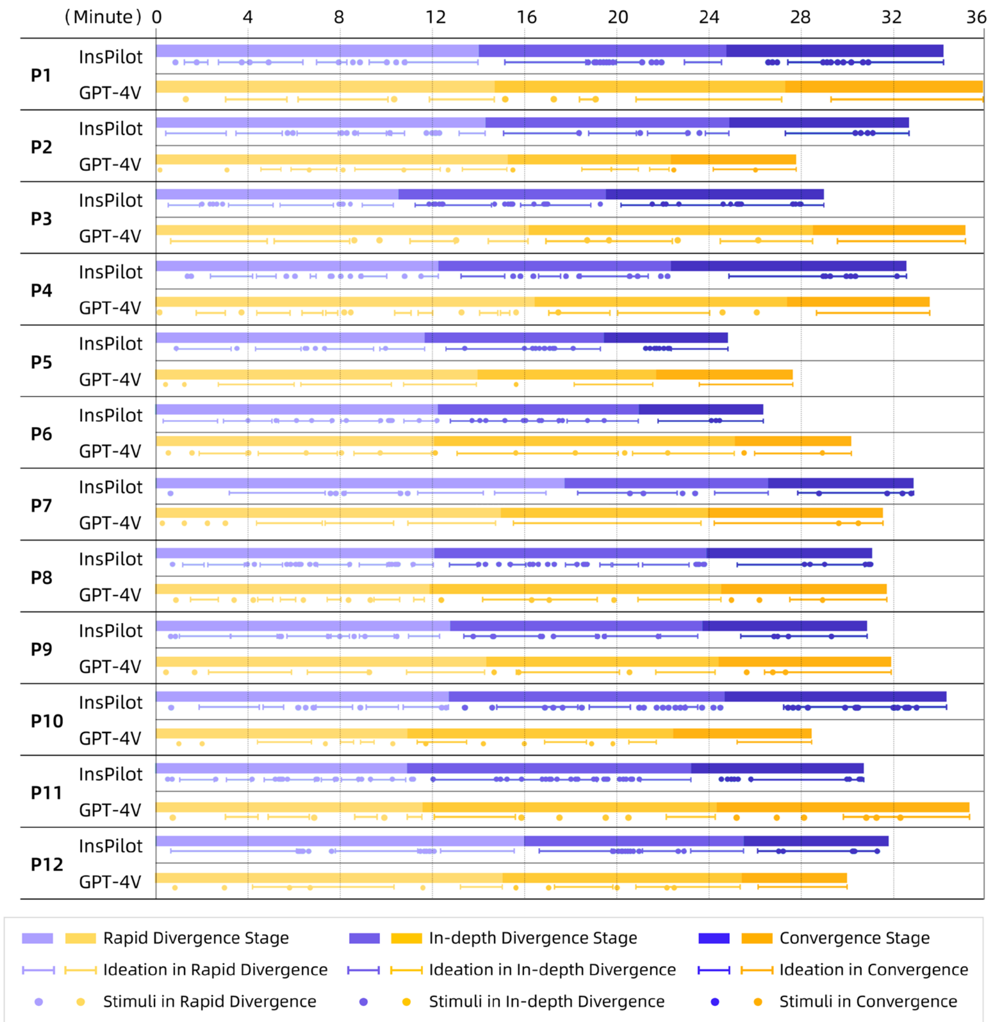

Figure A16. Results of the behaviour coding.