

DETC2025-XXXXXX

## HUMAN-GEN AI CO-DESIGN: EXPLORING FACTORS IMPACTING TRUST CALIBRATION

Chenjun Guo  
University of California  
Berkeley, CA

Antoni Borghini  
Université de Liège  
Liège, Belgium

Kosa Goucher-Lambert  
University of California  
Berkeley, CA

Gaëlle Baudoux  
University of California  
Berkeley, CA

### ABSTRACT

*The process of generating ideas during co-design with a Generative AI (GenAI) system requires the gradual calibration of trust in that system. Trust plays a pivotal role in shaping human interactions with technology, and developing well-calibrated trust is essential for the effective use and integration of GenAI. Proper trust calibration helps prevent underutilization of the system's capabilities and dissatisfaction with its output. For engineers and system designers, trust is particularly important as it directly influences user responses, system adoption, and overall engagement with new technologies. To explore the factors that influence trust fluctuation when co-designing with a GenAI system, we analyzed 12 hours of conceptual human-AI co-design sessions using a custom GenAI system capable of producing images across various generation modes from convergent-divergent to abstract-concrete, and combining text and sketch prompting. Focussing on each moment of interaction with GenAI-generated images, we conducted an incremental and qualitative coding of each trust-related extract from think-aloud protocols. Through this approach, we identified 23 key factors that cause fluctuations in trust. Our findings reveal a complex network of factors that impact trust calibration, offering insights into how GenAI systems can be designed to facilitate faster and more effective trust-building in human-GenAI collaborations.*

Keywords: Conceptual design, Human-AI collaboration, Generative AI, Trust calibration, Human factors.

### 1. INTRODUCTION

Idea generation is a fundamental stage in the design process [1] as it is critical in determining the performance of the final artefact [2]. In recent years, the advent of GenAI

systems has provided designers with new avenues for enhancing the idea generation activity. These GenAI systems facilitate brainstorming and concept generation [3], stimulate creativity [4, 5], and expand design exploration [6]. As GenAI becomes more integrated into creative workflows, designers' trust in these systems has emerged as a crucial factor, influencing both the effective use of AI systems and the integration of AI-generated outputs into the design process.

Trust is defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [7, p. 51]. In the frame of creativity tasks, a well-calibrated trust reflects an accurate understanding of the AI's capabilities, enabling designers to form an accurate mental model that guides their expectations. Without proper calibration, designers may experience undertrust, also called disuse—AI's potential is underestimated, leading to missed usage opportunities—or overtrust, also called abuse—AI's capabilities are overestimated, resulting in unrealistic expectations and deceiving results [8]. Therefore, understanding how designers adjust their mental models and calibrate their trust in these GenAI systems is essential for optimizing the collaboration between humans and GenAI.

This paper investigates the factors that influence trust calibration during interactions with a GenAI system in the context of conceptual human-AI co-design. This research seeks to answer the following question: *What factors contribute to fluctuations in trust during human-GenAI collaboration, and how can these factors be leveraged to improve trust calibration in the context of Generative AI-assisted design?*

### 2. RELATED WORK

#### 2.1 Trust calibration process

Lee and See [7, p.51] define trust as “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability”. They classify the basis of trust into three factors: (1) performance-based factors, which concern task efficiency, (2) process-based factors, which focus on consistency, and (3) purpose-based factors, where the system is perceived as honest and with benevolence motivation. Our work assumes that purpose-based trust factors are given, focusing instead on performance- and process-based elements that impact trust calibration. Previous research further distinguishes three categories of trust development: (1) dispositional trust, linked to individual’s general tendency to trust, (2) situational trust, that arises from the particular task context, and (3) learned trust, that develops while learning and using the system [8]. In this study, we control for participant backgrounds (profession and experience) and standardize the environment and tasks, as we primarily focus on learned trust.

Trust calibration processes differ across human-human, human-automation, and human-AI interactions, shaped by the relationship and context of the interaction [9–12]. Human-human trust tends to build gradually, as individuals demonstrate trustworthiness through their behaviors [9]. Human-automation trust often begins with an overtrust, that then quickly declines if the system fails [10]. Human-AI trust combines both patterns: potentially showing early overtrust and then evolving as the AI meets user expectations [12, 11].

GenAI introduces unique factors in trust calibration. Unlike deterministic AI systems, GenAI relies on stochastic processes, producing varied outputs from the same input due to random seed-based sampling [13]. This inherent variability leads to fluctuating trust levels, as users encounter different outputs for identical prompts. Moreover, design tasks are often ill-defined and lack a singular optimizable solution [14], requiring evaluative criteria that go beyond performance to include aesthetic alignment, contextual relevance, and accuracy in meeting personal intent. These factors make trust in GenAI more susceptible to individual preferences and creative expectations. Additionally, the iterative and exploratory nature of design work means trust in GenAI is continually reassessed as designers refine their ideas across multiple iterations. This cyclical trust calibration process, where expectations evolve through repeated interactions with the AI, contrasts with the more linear trust development seen in other domains [15–18].

While prior work has explored trust calibration in human-AI collaboration [12, 17, 18], there remains a gap in understanding how trust fluctuates when designers interact with GenAI systems. Our study aims to fill this gap by investigating trust factors during GenAI-assisted design.

## 2.2 Generative AI for design process

GenAI systems have found broad application in the early stages of design, enhancing ideation and fostering creativity across various fields, including architecture [3, 6], engineering design [19], product design [20], interior design [21], and graphic design [22]. Studies show that these systems enable

faster and more extensive design exploration [23, 24]. Beyan and Rossy [25] further demonstrate that GenAI systems facilitate both abstract thinking and tangible results, allowing users to transcend the limitations of realism and the physical world, thus enabling the exploration of novel concepts [26]. In addition, GenAI has been integrated into the entire design production pipeline, with systems like those developed by Yonder [27] and Li [28] supporting the generation of floor plans, 3D forms, and facade renderings from text input, streamlining the design workflow, from ideation to production.

Several GenAI models, including text-to-image, sketch-to-image, and hybrid systems, have been developed to support designers. Text-to-image systems like Midjourney, Stable Diffusion, and DALL-E are widely used in creative workflows. However, the literature points to several limitations: (1) the need for users to interrupt their workflow to generate images, (2) challenges in prompt engineering to achieve accurate results, (3) the desire for more controllable and precise images, and (4) the need for a more agile interaction with the system [3, 24, 25, 29, 6]. These issues, while technical, also relate to the challenges in engineering trust calibration within GenAI systems, as they affect user behavior and efficiency.

As GenAI evolves, sketch-to-image and hybrid systems have emerged to address some of these limitations. For example, Zhang et al. [29] developed systems enabling architectural designers to generate renderings from sketches, providing greater flexibility and control. Other advancements, like Gao’s [30] tailored system, integrate design knowledge and allow for highly specific outputs from intuitive sketches. Hybrid systems, such as Sketch2Prototype [31] and Bologan [32], combine both text and visual inputs, allowing designers to incorporate non-verbal thinking into the design process.

Despite the growing use of GenAI in design workflows, existing studies largely focus on the functionality of these systems and their impact on creativity, with limited attention given to how trust fluctuates during interactions with GenAI [24–26, 33]. The limitations identified in these systems—such as the need for prompt engineering and the challenge of maintaining a consistent workflow—are closely linked to trust calibration, yet few studies explore how designers adjust their trust when these challenges arise. Our study addresses this gap by examining how these limitations affect trust calibration in GenAI-assisted design. Specifically, we focus on a GenAI prototype that integrates both sketch and text inputs, offering enhanced flexibility and introducing new trust-related challenges. This approach provides valuable insights into improving human-GenAI collaboration and ensuring effective trust calibration throughout the design process.

## 3. METHODS

To explore the factors influencing trust fluctuations during interactions with a GenAI system, we observe architects in the conceptual design phase, co-designing with a custom-developed GenAI that integrates both sketch and text inputs, specifically created to study GenAI usage.

### 3.1 Design task

Architects were tasked with designing an in-law suite within a separate dwelling of an existing house. They were provided with a list of expected functional spaces (bedroom, bathroom, clothing space, and desk space), as well as a target surface area of less than 20 square meters. The experiment's task is subdivided into 4 successive sub-tasks explicitly delimited by a timer sound making the participant move to the following sub-task: (1) 10 minutes to explore the GenAI system interface and familiarize themselves with its functions while engaging in a thinking-aloud exercise to break the ice; (2) 20 minutes focusing on the functional design (zoning and layout of the functional spaces); (3) 20 minutes dedicated to form design (volumetry, façade, and materiality); and (4) 10 minutes of free design time to revisit function-related aspects if necessary, try a final AI generation, and/or synthesize the design proposition.

Choosing this architectural design task offers a unique opportunity to explore trust calibration in a complex and multifaceted context. Unlike engineering design tasks, which are often more objective and performance-driven, architectural design is inherently subjective and creative. It requires designers to make critical decisions about functionality, aesthetics, and spatial efficiency under tight constraints, while also navigating outputs that may be unexpected or misaligned with their creative intent—factors that can trigger fluctuations in trust. The open-ended nature of architectural design, with its trade-offs and iterative exploration, provides a rich environment for studying how designers adjust their expectations of AI. These elements introduce a more nuanced and dynamic context for investigating trust calibration in GenAI interactions. Additionally, the principles of design cognition and behavior are shared across disciplines, enabling us to derive valuable insights for engineering and system design.

### 3.2 Participant population

The study involved 12 graduate students in their final year of an architectural engineering program, each with a level of expertise comparable to professional designers in relation to the design task. In declaring their prior experience with GenAI systems, 3 individuals identified as novices and 9 as having limited experience (having experimented with GenAI but not using it regularly). The group consisted of 6 males, 6 females, and no individuals identifying as non-binary, with ages ranging from 21 to 25 years old. Although the study involved 12 participants, each contributed one hour of rich dynamic task data, including extensive think-aloud comments and video recorded behavior, which will be systematically hand-coded or detailed analysis. This in-depth data from each participant ensures a rich and sufficient dataset for our analysis of GenAI trust calibration.

### 3.3 Generative AI system

The system used for co-design in this experiment was developed internally to support various research initiatives on GenAI usage. Detailed descriptions of the system have been provided in a previous paper [34]. This system includes an

interface that supports GenAI-assisted co-design, integrating various models to produce images across different modalities.

The system operates on a tablet with a digital pencil and a bluetooth keyboard. The interface of the system (Figure 1) presents a sketching space on the right half of the screen, supporting fundamental sketching functionalities. This sketching space also serves as the input prompt when prompting image generation in sketch mode. The central column displays additional prompting parameters: specifying the type of representation sketched (facade - interior - floor plan) to help the model understand, specifying the desired generation type (rendering the sketched idea - inspire an alternative idea based on the sketch and on a chosen reference image), and specifying the desired image output (realistic for concrete proposition- sketched for more abstract proposition). On the top of that central column, the user can switch to a text input mode, in which a text box opens and allows the user to enter a textual prompt.

A “Generate Image” button triggers the image generation process. On the left side of the screen, the system displays three generated images, from which the designer can either choose to trash, add to the project mood board, or let live in the generation library. Both the generation library and the project mood board can be accessed via buttons on the top left corner of the screen.

The AI models used for image generation were selected for their high performance in architectural design tasks. Information about the design brief is pre-fed into the prompting architecture to ensure that the generated images align with the design topic and scale. Fixed generation instructions further ensure that key elements (such as material, furniture, and style) are accurately translated into the generated images.

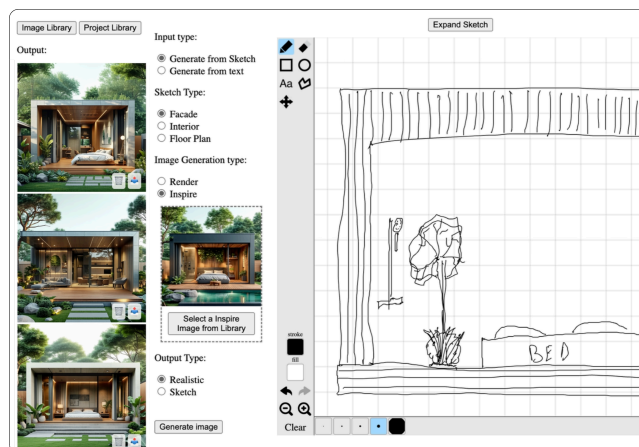


FIGURE 1: INTERFACE OF THE GEN AI SYSTEM USED

### 3.4 System's capacity

The system employed ChatGPT 4.0 for sketch interpretation, incorporating fixed instructions related to the design brief, and then generated images using Stability (SD3 Large). Further details on the system architecture can be found

in our earlier work [34]. Preliminary testing demonstrated the system's strong capability in producing renderings aligned with the design brief. Figure 2 illustrates outputs from our study.



FIGURE 2: ILLUSTRATION OF THE SYSTEM'S CAPACITY

### 3.5 Data collection

Data was collected through three channels: (1) verbal think-aloud protocols, where participants articulated their design rationale, intentions, and any fluctuations in trust as they interacted with the GenAI system ; (2) a camera recording the design session to capture a backup record of the experiment and provide context for interpreting all collected data, including any disruptions; and (3) an interface log that recorded every action taken by the participants, such as generating new images, using text or sketch-based prompts, selecting inspiration images, exporting images to the mood board, and deleting images, with corresponding time stamps. In total, 12 hours of Human-GenAI co-design were recorded, capturing 144 image generation interactions and 896 trust-related commentary instances.

### 3.6 Data coding

The primary objective of the analysis is to identify factors contributing to fluctuations in designers' trust during their interactions with the GenAI system . We filtered out all user actions that could potentially indicate trust increases or decreases, focusing on the following actions:

- generating new images using text-based or sketch-based prompts (with or without an inspiration image);
- exporting an image to the library (indicating satisfaction);
- deleting an image (indicating dissatisfaction);
- and modifying a sketch or text prompt to regenerate an image.

For each action, we analyzed the think-aloud to determine:

- whether the generated image aligned with the user's mental model;
- if trust fluctuated (increased or decreased) as a result;
- and which specific element of the gap between the user's expectations and the generated image contributed to the change in trust.

We applied incremental qualitative coding to categorize the aspects of expectation gaps into specific factors. Once all instances of trust fluctuation were identified (both increases and decreases, along with the corresponding factors), we grouped these factors into broader categories of determinants that influence trust in GenAI.

Figure 3 illustrates the coding process. At timestamp 09:35, the user performed a relevant action (image generation), prompting us to record think-aloud data. We extracted comments on the generated images, focusing on deviations from user expectations. These differences were then categorized into factors, sticky notes in the figure.






Timestamp	9:36 (Action: Generate)				
Input	 				
Output	  				
Think Aloud	<p>-1 not bad at all, respected my goal</p> <p>-2 not successful, but the <u>glasses and bay windows</u> are there</p> <p>-3: not successful, not library anymore, and as for form, i did not go crazy, it follows my drawing (rectangle), but did not surprise me</p>				
Factors	- Conceptual Integrity	+ Conceptual Integrity	- Element Fidelity	+ Symbol Interpretation	- Inspirational Value

FIGURE 3: DATA CODING EXAMPLE



### 3.7 Measures

For each factor identified (Section 3.6), we count the number of participants which mentioned this factor, the number of comments of this factor which were mentioned by each participant, and count the number of positive and negative comments separately. The following parameters are defined to access the detailed properties of these factors.

#### 3.5.1 Sentiment Direction (SD)

Sentiment Direction (SD) quantifies the extent to which a factor is associated with trust increase or decrease. The scale of SD is  $[-1,1]$ , where a positive value indicates a stronger association with trust increase, with higher scores reflecting a greater positive sentiment, and vice versa for negative values. SD is computed as the average of individual participant sentiment scores ( $SD_i(f)$ ), which normalizes sentiment by considering the proportion of positive and negative comments within the factor, ensuring that a higher presence of positive comments results in a positive score, while a higher presence of negative comments results in a negative score.

For each participant  $i$  on factor  $f$ :

$$SD_i(f) = \left( \frac{P_i(f) - N_i(f)}{P_i(f) + N_i(f)} \right) \quad (1)$$

For factor  $f$ :

$$SD(f) = \frac{1}{N} \left( \sum_{i=1}^N SD_i(f) \right) \quad (2)$$

Where:

$SD_i(f)$ : Individual participant sentiment score on factor  $f$ .

$P_i(f)$ : The positive comments on factor  $f$  by participant  $i$ .

$N_i(f)$ : The negative comments on factor  $f$  by participant  $i$ .

#### 3.5.2 Sentiment Agreement (SA)

Sentiment Agreement (SA) quantifies the degree of variability in participants' sentiment toward a given factor. A high SA indicates strong divergence in opinions, suggesting that participants have conflicting views on whether the factor contributes to trust increase or decrease. Conversely, a low SA suggests general consensus, indicating that most participants share a similar sentiment regarding the factor. The scale of SA is  $[0,+\infty)$ . SA is computed as the standard deviation of individual participant sentiment scores ( $SD_i$ ).

For factor  $f$ :

$$SA(f) = \sqrt{\frac{1}{N} \sum_{i=1}^N (SD_i(f) - SD(f))^2} \quad (3)$$

#### 3.5.3 Importance Agreement (IA)

Importance Agreement (IA) quantifies the degree of variation in how much attention participants allocate to a given factor. A high IA indicates that some participants place significant emphasis on the factor while others barely mention it, suggesting an uneven distribution of importance. Conversely, a low IA suggests that participants consider the factor with similar levels of attention, indicating its consistent relevance across users. The scale of IA is  $[0,+\infty)$ . IA is computed as the standard deviation of the proportion of comments participants dedicates to the factor ( $I_i(f)$ ), relative to their total comments.

For each participant  $i$  on factor  $f$ :

$$I_i(f) = \frac{\text{Total comments by participant } i \text{ on factor } f}{\text{Total comments by participant } i} \quad (4)$$

For factor  $f$ :

$$IA(f) = \sqrt{\frac{1}{N} \sum_{i=1}^N (I_i(f) - I_{\text{mean}}(f))^2} \quad (5)$$

Where:

$I_i(f)$ : Proportion of participant  $i$ 's total comments that are about factor  $f$ .

## 4. RESULTS

This section first presents the key factors influencing trust fluctuations, categorized into six thematic categories based on the aspects of the system's capabilities they reflect. We then analyze, category by category, the factors showing higher impact on trust fluctuation, using Sentiment Direction (SD) and Sentiment Agreement (SA) measures, and the factors more frequently mentioned by the users, in terms of their total occurrence count and Importance Agreement (IA).

### 4.1 Factors Influencing Trust Fluctuation

Answering the first part of our research question, "*What factors contribute to fluctuations in trust during human-GenAI collaboration?*", we firstly identified an extensive list of factors that influenced trust. Through further analysis and synthesis, we narrowed these down into six broad categories, each representing a distinct dimension of system capabilities that inform user trust calibration (Table 1).

The **Intent Alignment** category captures the GenAI's ability to interpret user input correctly and maintain the intended design intent. The **Design exploration** category relates to the GenAI's capacity to expand design possibilities and encourage creative exploration. The **Aesthetics** category focuses on the visual and perceptual qualities of GenAI-generated outputs. The **Plausibility** category assesses the realism and feasibility of GenAI-generated outputs. The **Iterative Adaptability** category evaluates the GenAI's responsiveness to iterative modifications. Finally, the **Consistency** category addresses the GenAI's ability to maintain stability in its generated outputs.

**TABLE 1: TRUST FLUCTUATION CATEGORIES AND RESPECTIVE FACTORS.**

Intent Alignment		Design Exploration	
<i>Symbol Interpretation</i>	Correctly understanding symbols and incorporating the corresponding intention	<i>Unexpected Design Directions</i>	Generating unexpected or unmentioned ideas
<i>Completeness</i>	Meeting all the requirements in the prompt	<i>Generation Diversity</i>	Generating a variety of outputs
<i>Composition Adherence</i>	Maintaining spatial arrangements	<i>Inspirational Value</i>	Generating outputs that spark new ideas
<i>Conceptual Integrity</i>	Preserving the “spirit” of the design	<i>System Fixation</i>	Tendency of repeating elements or styles
<i>Element Fidelity</i>	Accurately incorporating key elements and their features from the sketch	<i>Alternative Solutions</i>	Providing alternative solutions for the same function
<i>Scale Accuracy</i>	Ensuring design elements have appropriate size and volume	<b>Aesthetics</b>	
<i>Viewpoint Selection</i>	Choosing a viewpoint that displays the key elements	<i>Material Accuracy</i>	Material choices aligning with design intent.
<i>Spontaneous Components</i>	Objects appearing without logical reasoning cohesive to the intent	<i>Atmosphere Match</i>	Matching emotional or immersive abstract quality for the generated space
<i>Implicit Property Interpretation</i>	Correctly inferring and incorporating implicit properties independently of what is sketched	<i>Aesthetic Quality</i>	Ensuring the visual appeal and artistic coherence of the generation
Plausibility		Iterative Adaptability	
<i>Constructability</i>	Proposing a design that is structurally feasible for real-world construction	<i>Iterative Adaptability</i>	Refining and evolving previous generations based on user input
<i>Functional Usability</i>	Ensuring that generated elements serve their intended practical function	<b>Consistency</b>	
<i>Adherence to Design Conventions</i>	Aligning with established architectural principles and design norms	<i>Coherence Consistency</i>	Sustaining a stable level of design coherence across multiple generations

## 4.2 Analysis of Factors in Relation to Occurrence and Sentiment Direction

Figure 4 illustrates each factors’ Sentiment Direction (SD), Sentiment Agreement (SA), Average Occurrence, and Importance Agreement (IA).

### 4.2.1 Intent Alignment

Four factors in this category—*Conceptual Integrity*, *Completeness*, *Element Fidelity*, and *Symbol Interpretation*—are among those with the highest occurrence.

*Conceptual Integrity* and *Completeness* exhibited both high occurrence and high IA values, indicating variability in how frequently participants commented on these factors, with some referencing them consistently while others rarely mentioned them. Observations from the user study suggest that participants who comment on *Conceptual Integrity* frequently tended to mention *Completeness* less, and vice versa. Participants who emphasized on *Conceptual Integrity* had a strong guiding idea or thematic direction they wanted the system to preserve, expressing positive sentiment when it

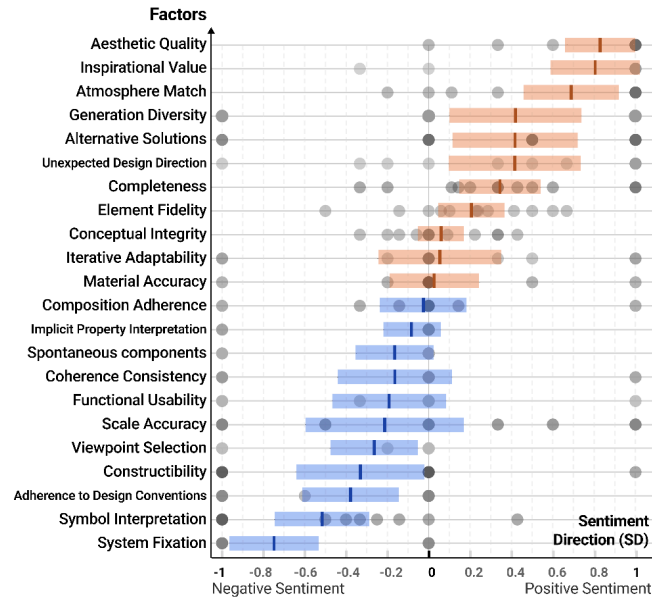
captured their vision and negative sentiment when it deviated. In contrast, those focused on *Completeness* sought to include a comprehensive set of design elements in a single generation, responding positively when most components were present. Both *Conceptual Integrity* and *Completeness* have a relative neutral sentiment direction.

*Element Fidelity* was another frequently occurring factor, with participants expressing positive sentiment when the system accurately integrated key design elements from their input and negative sentiment when it failed to do so. This factor also had a slightly positive but mostly neutral sentiment direction, as instances where participants observed expected elements were nearly as frequent as those where missing elements led to trust decrease.

*Symbol Interpretation* is among both the highest occurrence factors and those associated with the most negative sentiment. Participants relied on universal symbols (e.g., desk, window, wardrobe) to convey design intent, with trust decreasing when the GenAI failed to interpret and incorporate these symbols accurately. This factor exhibited high SA,

reflecting mixed experiences—some participants found the system effectively aligned with their symbolic input, while others encountered frequent misinterpretations.

The remaining factors in this category, *Composition Adherence* and *Scale Accuracy*, exhibited moderate occurrence and a neutral sentiment direction. *Scale Accuracy* had a high SA, as some participants considered the size of elements crucial, while others paid little attention to it.



**FIGURE 4: FACTORS' SENTIMENT DIRECTION, SENTIMENT AGREEMENT, OCCURRENCE AND IMPORTANCE AGREEMENT**

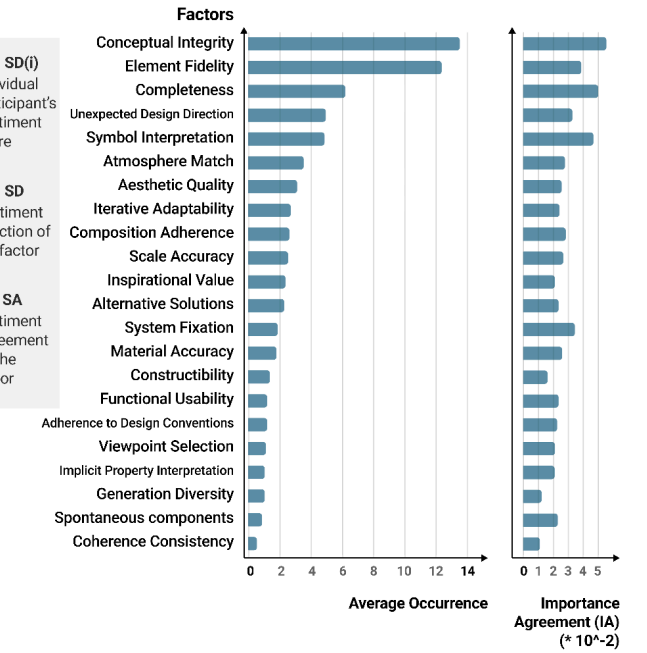
#### 4.2.2 Design Exploration

This category includes factors exhibiting highly polarized sentiment directions. *Unexpected Design Direction*, *Inspirational Value*, *Alternative Solution*, and *Generation Diversity* have positive SD.

*Inspirational Value* increased trust when the system stayed within the participant's intended design direction but introduced new ideas which implement the original ones. For instance, one participant, aiming to generate a rounded roof, noted how "it gives me inspiration to push the form to the limit—imagine a dome-shaped facade" (P4).

*Unexpected Design Direction* increased trust when participants observed the GenAI introducing relevant yet unconsidered elements. For instance, when aiming to design a fancy terrace, the system generated a set of outdoor lightings not mentioned in the sketch. *Alternative Solutions* increased trust by presenting different methods that still addressed the intended design problem. For instance (Figure 5), rather than producing the requested study desk and chairs, the system generated a sofa and side table, which served a similar function of supporting reading activities in the library. Generation Diversity boosted trust by offering a range of diverse but

*Viewpoint Selection*, *Spontaneous Components*, and *Implicit Property Interpretation* were less frequent and typically relevant to only specific design intents. For example, *Viewpoint Selection* was important when the perspective influenced the sketch's presentation. *Implicit Property Interpretation* mainly occurred with participants perceiving human-like qualities in the system (discussed in 5.4.2) and became more common in the later stages of the experiment



coherent design solutions aligned with participants' intentions in one generation. However, *Unexpected Design Direction*, *Alternative Solution* and *Generation Diversity* all exhibited high SA values, reflecting diverse sentiment—some viewed unexpected elements as beneficial, while others perceived them as unintentional deviations.



**FIGURE 5: EXAMPLE OF TRUST INCREASE DUE TO ALTERNATIVE SOLUTION**

In contrast, *System Fixation* displayed a strong negative SD, reducing trust when the system repeatedly generated similar components with limited variation. Its moderate SSD

and occurrence suggest that participants largely agreed on its negative impact, though it was not an uncommon phenomenon.

#### 4.2.3 Aesthetics

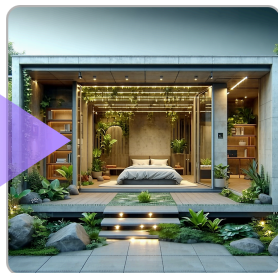
*Aesthetic Quality* and *Atmosphere Match* were strongly associated with trust increases, often highlighted when GenAI-generated outputs offered significant visual appeal or successfully captured participants' intended ambiance. *Aesthetic Quality* had a low SA, indicating broad agreement on its positive impact. However, *Atmosphere Match* showed a moderate SA, reflecting varied opinions. While many participants appreciated the generated atmosphere, some found it unsuitable for their design goals.

*Material Accuracy* appears a neutral sentiment. It leads to trust increase when participants find that the material applied in the generation fits in the design intent and atmosphere; however, people also criticize it when their prompted material was not presented or didn't reach their expectation.

#### 4.2.4 Plausibility

The factors in this category primarily reduce trust. *Adherence to Design Conventions* and *Functional Usability* led to trust decrease when participants found GenAI outputs inconsistent with standard practices or functionally impractical. Both exhibited moderate-to-high SA, indicating that participants mostly commented on design failures while rarely acknowledging successful adherence or functionality. For example (Figure 6), one participant requested an atrium in the middle of a house but received an indoor atrium, which typically conflicts with common architectural norms. *Constructability* had a mildly negative sentiment, as participants often criticized designs deemed unrealistic yet occasionally praised those that appeared industry-ready. All three factors had a middle-to-low occurrence overall.

A square pavilion that has an atrium in the middle, this atrium is the garden. Around the atrium are the open spaces and the bedroom. The bedroom is separated from the open space.



**FIGURE 6:** EXAMPLE OF TRUST DECREASE DUE TO POOR ADHERENCE TO DESIGN CONVENTIONS

#### 4.2.5 Consistency

*Coherence Consistency* often led to trust decreases when participants observed inconsistencies across generations, making the GenAI appear unreliable. One participant remarked, "It is crazy that it generated a tree in the middle of the house before, but now it cannot reproduce one" (P10). This factor had relatively low occurrence.

#### 4.2.6 Iterative Adaptability

*Iterative Adaptability* showed a mild sentiment (SD) and moderate occurrence, reflecting its relationship to human-like interaction, which is further discussed in Section 5.4.2.

### 5. DISCUSSION

This section explores explanations for the observed variations in sentiment and factor occurrence, offering insights into why certain factors influence more significantly trust fluctuations. Based on our findings, we propose design principles for improving GenAI systems in conceptual design. Finally, we discuss the impact of perceived human-like qualities on trust calibration.

#### 5.1 Hypothesizing the rationale behind the results

Sections 4.2 and 4.3 respectively identified the factors with the most polarized sentiment and highest occurrence. In this section, we hypothesize the underlying reasons for these trends and propose design principles for improving GenAI systems for conceptual design.

We posit that *Aesthetic Quality* and *Atmosphere Match* are more strongly associated with positive sentiment due to two key reasons. First, Stability AI is trained on publicly available online images, many of which are curated for showcasing pleasing aesthetics as talked before. This likely biases the AI towards generating highly polished and visually appealing outputs. Second, designers are inherently visual thinkers, trained to be highly sensitive to aesthetic qualities and spatial ambiance. As a result, they exhibit heightened positive responses when the GenAI produces outputs that are visually striking or align with the intended atmosphere of their design concept, which in turn amplifies the frequency of these factors being mentioned.

Another set of factors with high positive sentiment includes *Inspirational Value*, *Unexpected Design Direction*, and *Generation Diversity*. The design brief stated that the primary goal of the session was to explore ideas and seek inspiration for a design task. As a result, we hypothesize that participants were open to new design possibilities. The strong positive sentiment associated with these factors supports the notion that GenAI can effectively stimulate creativity in design. However, the high SSD score indicates significant variation in how participants perceived inspiration. This suggests that the boundary between an inspiring suggestion and an unrelated or irrelevant generation is inherently subjective. While some participants embraced unexpected outputs as sources of inspiration, others perceived them as deviations from their intended design goals, leading to more polarized reactions.

Factors highly associated with negative sentiment include *Adherence to Design Conventions*, *Functional Usability*, and *Coherence Consistency*, which were frequently mentioned when participants encountered "weird images". We hypothesize that participants generally do not comment on realistic or feasible generations when such qualities are consistently met. Instead, they tend to verbalize their thoughts only when a generation appears unrealistic, impractical, or inconsistent with

design norms. This selective attention likely explains why these factors are predominantly linked to negative sentiment, as they are more noticeable in cases where the GenAI fails to meet expected design standards rather than when it performs as anticipated. The reason why *Symbol Interpretation* is associated with negative sentiment will be further discussed in Section 5.4.

The three most frequently occurring factors—*Conceptual Integrity*, *Completeness*, and *Element Fidelity*—all belong to the Intent Alignment category. We hypothesize that this is because understanding whether the GenAI correctly interprets user intent and whether it can generate outputs that align with the designer's expectations is the primary concern when adopting a GenAI system. Moreover, these factors are also among the most immediate and straightforward to evaluate. Unlike factors such as *Aesthetic Quality*, which may not be a focus in every generation, *Conceptual Integrity*, *Completeness*, and *Element Fidelity* can be assessed consistently across all outputs. For many participants, these factors likely serve as a baseline expectation when engaging with the GenAI, forming the foundation upon which trust in the system is established.

## 5.2 Design principles for Generative AI for design inspiration

A clear pattern emerges from the analysis: factors from the Intent Alignment category emerged as the most impactful, while those linked to positive sentiment primarily stemmed from Design Exploration and Aesthetics categories, and negative sentiment was most often associated with the Plausibility category. Based on these findings, we propose the following principles for improving user experience and design inspiration when designing new GenAI systems.

**Prioritize intent alignment:** Ensure the system accurately interprets and reflects the designer's input, particularly in terms of conceptual integrity. This fosters trust by making the generated outputs relevant, aligned, with the designer's vision.

**Improve completeness:** Ensure the system generates both focal elements and supporting components of a design in a coherent manner, promoting a more integrated design process.

**Enhance aesthetic quality:** Focus on creating outputs that align with the desired ambience and exhibit high aesthetic appeal. This positively impacts user creative exploration.

**Ensure plausibility:** Align the system's outputs with common design conventions, verify feasibility, and maintain coherence across iterations. This helps mitigate negative reactions and enhances user confidence in the system's outputs.

This answers our second research question “*How can these factors be leveraged to improve trust calibration in the context of GenAI-assisted design?*”

## 5.3 Participants' perception of human-like factors and their impact on trust calibration

During the user study, participants frequently attributed human-like characteristics to the system. 10 out of 12 participants made remarks such as “*It likes this idea*”, “*It really likes wooden doors*”, or “*It respects my composition a lot*”. These observations suggest that certain system behaviors led

participants to perceive the GenAI as exhibiting preferences, or intent, similar to human interaction.

### 5.4.1 Factors contributing to human-like perception in trust calibration

We hypothesize that the **randomness in generation quality**—across categories such as Plausibility, Intent Alignment, Aesthetics, and Design Exploration—leads designers to perceive the GenAI as more human-like, mirroring human communication rather than a deterministic machine. This variability in generation quality is indeed alike real-world design negotiations, where differences in interpretation arise between designers and stakeholders. Participants seemed to expect the GenAI to process sketches similarly to a human.

For example, in Figure 7, a participant intended to place a tree inside a house, but the system exhibited variability in its response, generating one facade with no tree, another with a tree behind a window, and a third with a tree integrated into the facade, stylized with a hat-like shape. This inconsistency resembled the way humans might interpret the same conceptual input differently, reinforcing the perception of the GenAI as an adaptive and interpretive agent rather than a rigid system.



FIGURE 7: HUMAN-LIKE INTERPRETATION RANDOMNESS

Another factor contributing to the human-like perception of the GenAI was **symbol interpretation**. Participants developed greater trust in the GenAI when they observed that it could correctly interpret abstract symbols commonly used in design workflows—an ability typically associated with human designers. For instance, in Figure 8, a participant used blue outlines to indicate glass and diagonal lines to represent reflective lighting. The GenAI successfully incorporated these symbolic cues into the rendering, reinforcing the participant's perception of human-like interaction in the design process.

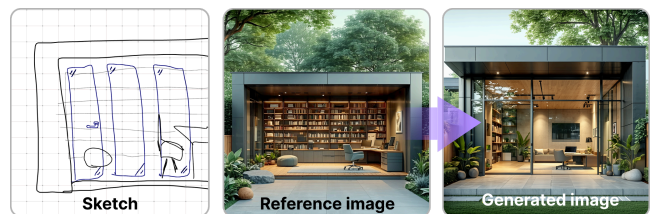


FIGURE 8: HUMAN-LIKE SYMBOL INTERPRETATION



#### 5.4.2 Impact of human-like interaction on trust calibration

We observed that participants generally experienced a change in communication strategy when they first perceived human-like interactions with the system. This perception led them to adopt communication approaches similar to those used with human designers. In Lee and See's theory [7], trust between humans and machines tends to decline immediately when imperfections are noticed. In contrast, trust in human-to-human relationships erodes gradually over multiple instances of unmet expectations [9]. We observed that the perception of human-like interaction in GenAI appears to shift the trust calibration process closer to human-to-human dynamics, where participants exhibited more patience and adaptability in refining their interactions with the system, which results in a **longer calibration process**, as trust does not immediately decline after a single unsuccessful generation.

For example, in Figure 9, a participant intended to use a closet or cabinet to separate the bedroom and bathroom areas. In the first two rounds, the generated outputs did not align with her expectations. However, instead of disengaging, she adjusted her explanation, and in the third round, the first generated image successfully reflected her intent. This iterative process suggests that the perception of human-like interaction encouraged participants to persist in refining their prompts rather than dismissing GenAI's capabilities after initial failures

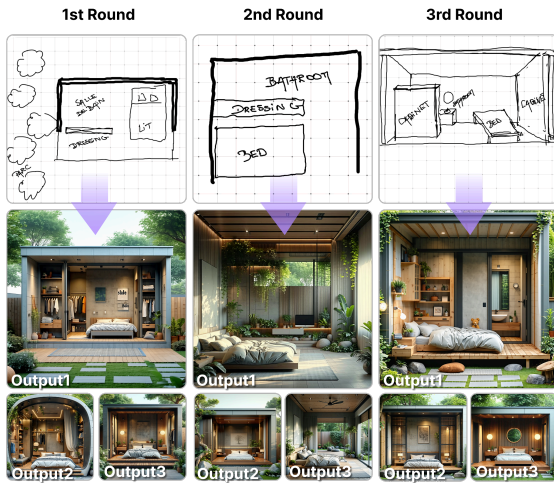


FIGURE 9: LENGTHENED TRUST CALIBRATION PROCESS

Additionally, we observed that when participants perceived human-like qualities in the GenAI, they began **expecting the system to interpret implicit meaning**, rather than providing exact sketches of their intent—a behavior similar to discussions with human designers. For instance, in Figure 10, a participant aimed to generate an atrium at the center of a house. After the first unsuccessful attempt, instead of assuming that the GenAI lacked the capability to generate an atrium, he interpreted the failure as the system hiding the atrium behind other elements. Consequently, he adjusted his prompts in the next two generations, emphasizing that the atrium should be clearly

visible. In the final iteration, although the atrium should not have been visible from the chosen viewpoint due to obstructing walls, he still included it in the sketch, believing that the system would understand his intent beyond literal representation, and in the end he finally got satisfying generations.

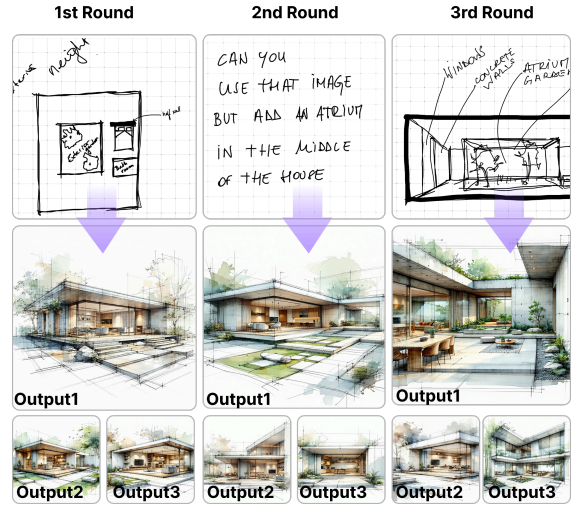


FIGURE 10: HUMAN-LIKE RESULTING EXPECTATION

The perception of human-like interaction also explains why the factor *Symbol Interpretation* is associated with negative sentiment. Initially, participants relied on universal symbols to communicate their design intent. However, as they began expecting the system to understand implicit meanings, their symbolic sketches became more complex and interpretative, increasing the difficulty of accurate interpretation—similar to the atrium case discussed earlier. As a result, participants experienced greater frustration when the system failed to infer their intent correctly, leading to negative sentiment due to less likelihood of achieving desired outcome.

#### 6. CONCLUSION

This study examined the factors influencing trust calibration in human-GenAI co-design, based on 12 hours of co-design sessions and 896 trust-related comments. We identified 23 key factors contributing to fluctuations in trust, highlighting key levers for improving trust calibration: prioritizing intent alignment, ensuring aesthetic quality, and addressing plausibility through adherence to design conventions, feasibility verification, and consistency. Additionally, the study revealed that variability in generation quality and effective symbol interpretation enhanced the perception of GenAI as more human-like, fostering longer calibration periods and greater patience with errors. These findings provide valuable and actionable insights for designing GenAI systems that better integrate with designers' workflows and build trust. While the study's limited sample size and focus on a single design case study constrain generalizability, future research should explore broader design contexts and the long-term impact of trust on GenAI system adoption.

## REFERENCES

- [1] Leahy, Keelin, Seifert, Colleen, Daly, Shanna and McKilligan, Seda. "Overcoming design fixation in idea generation." (2018).
- [2] Roberts, Matthew, Allen, Stephen and Coley, David. "Life cycle assessment in the building design process—A systematic literature review." *Building and Environment* Vol. 185 (2020): p. 107274.
- [3] Nagele, Julia. "Fantasy on Demand: The Temptation Of Text-to-Image AI." *CTBUH Journal* No. 3 (2023).
- [4] Karimi, Pegah, Rezwana, Jeba, Siddiqui, Safat, Maher, Mary Lou and Dehbozorgi, Nasrin. "Creative sketching partner: an analysis of human-AI co-creativity." *Proceedings of the 25th international conference on intelligent user interfaces*: pp. 221–230. 2020.
- [5] Kim, Jingoog, Maher, Mary Lou and Siddiqui, Safat. "Collaborative Ideation Partner: Design Ideation in Human-AI Co-creativity." *CHIRA*: pp. 123–130. 2021.
- [6] Paananen, Ville, Oppenlaender, Jonas and Visuri, Aku. "Using text-to-image generation for architectural design ideation." *International Journal of Architectural Computing* Vol. 22 No. 3 (2024): pp. 458–474.
- [7] Lee, John D and See, Katrina A. "Trust in automation: Designing for appropriate reliance." *Human factors* Vol. 46 No. 1 (2004): pp. 50–80.
- [8] Lotfalian Saremi, Mostaan and Bayrak, Alparslan Emrah. "A Survey of Important Factors in Human-Artificial Intelligence Trust for Engineering System Design." *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 85420: p. V006T06A056. 2021. American Society of Mechanical Engineers.
- [9] Rempel, John K, Holmes, John G and Zanna, Mark P. "Trust in close relationships." *Journal of personality and social psychology* Vol. 49 No. 1 (1985): p. 95.
- [10] Dzindolet, Mary T, Peterson, Scott A, Pomranky, Regina A, Pierce, Linda G and Beck, Hall P. "The role of trust in automation reliance." *International journal of human-computer studies* Vol. 58 No. 6 (2003): pp. 697–718.
- [11] de Visser, Ewart J, Krueger, Frank, McKnight, Patrick, Scheid, Steven, Smith, Melissa, Chalk, Stephanie and Parasuraman, Raja. "The world is not enough: Trust in cognitive agents." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 56. 1: pp. 263–267. 2012. Sage Publications Sage CA: Los Angeles, CA.
- [12] Lucas, Gale M, Becerik-Gerber, Burcin and Roll, Shawn C. "Calibrating workers' trust in intelligent automated systems." *Patterns* Vol. 5 No. 9 (2024).
- [13] Chen, L, Zhang, Zhe, Yang, Gang, Zhou, Qin, Xia, Yan and Jiang, Chao. "Evidence-theory-based reliability analysis from the perspective of focal element classification using deep learning approach." *Journal of Mechanical Design* Vol. 145 No. 7 (2023): p. 071702.
- [14] Li, Xingang, Wang, Ye and Sha, Zhenghui. "Deep learning methods of cross-modal tasks for conceptual design of product shapes: A review." *Journal of Mechanical Design* Vol. 145 No. 4 (2023).
- [15] Wynn, David C and Eckert, Claudia M. "Perspectives on iteration in design and development." *Research in Engineering Design* Vol. 28 (2017): pp. 153–184.
- [16] Khan, Shahroz, Kaklis, Panagiotis and Goucher-Lambert, Kosa. "How does agency impact human-AI collaborative design space exploration? A case study on ship design with deep generative models." *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 87318: p. V03BT03A055. 2023. American Society of Mechanical Engineers.
- [17] Chong, Leah, Raina, Ayush, Goucher-Lambert, Kosa, Kotovsky, Kenneth and Cagan, Jonathan. "The evolution and impact of human confidence in artificial intelligence and in themselves on AI-assisted decision-making in design." *Journal of Mechanical Design* Vol. 145 No. 3 (2023): p. 031401.
- [18] Chong, Leah, Zhang, Guanglu, Goucher-Lambert, Kosa, Kotovsky, Kenneth and Cagan, Jonathan. "Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice." *Computers in Human Behavior* Vol. 127 (2022): p. 107018.
- [19] Ma, Kevin, Moore, George, Shyam, Vikram, Villarrubia, James, Goucher-Lambert, Kosa and Reynolds Brubaker, Eric. "Human-AI Collaboration Among Engineering and Design Professionals: Three Strategies of Generative AI Use." *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 88407: p. V006T06A025. 2024. American Society of Mechanical Engineers.
- [20] Lee, Yu-Hsu and Chiu, Chun-Yao. "The impact of AI text-to-image generator on product styling design." *International Conference on Human-Computer Interaction*: pp. 502–515. 2023. Springer.
- [21] Chen, Junming, Shao, Zichun and Hu, Bin. "Generating interior design from text: A new diffusion model-based method for efficient creative design." *Buildings* Vol. 13 No. 7 (2023): p. 1861.
- [22] Vashishtha, Shanu, Prakash, Abhinav, Morishetti, Lalitesh, Nag, Kaushiki, Arora, Yokila, Kumar, Sushant and Achan, Kannan. "Chaining text-to-image and large language model: A novel approach for generating personalized e-commerce banners." *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*: pp. 5825–5835. 2024.
- [23] Jaruga-Rozdolska, Anna. "Artificial intelligence as part of future practices in the architect's work: MidJourney generative tool as part of a process of creating an architectural form." *Architectus* No. 3 (71 (2022): pp. 95–104.
- [24] Beyan, Eleonora Vilgia Putri, Rossy, Anastasya Gisela Cintya et al. "A review of AI image generator: influences, challenges, and future prospects for architectural field." *Journal of Artificial Intelligence in Architecture* Vol. 2 No. 1 (2023): pp. 53–65.
- [25] Beyan, Eleonora Vilgia Putri, Rossy, Anastasya Gisela Cintya et al. "A review of AI image generator: influences, challenges, and future prospects for architectural field."

*Journal of Artificial Intelligence in Architecture* Vol. 2 No. 1 (2023): pp. 53–65.

- [26] Casakin, Hernan and Wodehouse, Andrew. “A systematic review of design creativity in the architectural design studio.” *Buildings* Vol. 11 No. 1 (2021): p. 31.
- [27] Yonder, Veli Mustafa, Dulgeroglu, Ozum, Dogan, Fehmi and Cavka, Hasan Burak. “The Role of the Computational Designer From Computer-Aided Design To Machine Learning-Aided Design a Study on Generative Models and Design Prompts.” *41st Conference on Education and Research in Computer Aided Architectural Design in Europe (ECAADE)–SEP 18-23, 2023–Graz Univ Technol, Graz, AUSTRIA*. 2023. Ecaade-education & Research Computer Aided Architectural design Europe.
- [28] Li, Pengzhi and Li, Baijuan. “Generating daylight-driven architectural design via diffusion models.” *arXiv preprint arXiv:2404.13353* (2024).
- [29] Zhang, Chengzhi, Wang, Weijie, Pangaro, Paul, Martelaro, Nikolas and Byrne, Daragh. “Generative image AI using design sketches as input: Opportunities and challenges.” *Proceedings of the 15th Conference on Creativity and Cognition*: pp. 254–261. 2023.
- [30] Gao, Jin and Patel, Sayjel. “From Sketch to Design: A Cross-scale Workflow for Procedural Generative Urban Design.” *Proceedings of the 29th CAADRIA Conference*: pp. 343–352. 2024.
- [31] Edwards, Kristen M, Man, Brandon and Ahmed, Faez. “Sketch2Prototype: rapid conceptual design exploration and prototyping with generative AI.” *Proceedings of the Design Society* Vol. 4 (2024): pp. 1989–1998.
- [32] Bolojan, DANIEL, Vermisso, EMMANOUIL and Yousif, SHERMEEN. “Is language all we need? A query into architectural semantics using a multimodal generative workflow.” *POST-CARBON, Proceedings of the 27th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA)*, Vol. 1: pp. 353–362. 2022.
- [33] Elasri, Mohamed, Elharrouss, Omar, Al-Maadeed, Somaya and Tairi, Hamid. “Image generation: A review.” *Neural Processing Letters* Vol. 54 No. 5 (2022): pp. 4609–4646.
- [34] Baudoux, Gaelle and Goucher-Lambert, Kosa. “Multimodal generative AI for conceptual design: Enabling text-based and sketch-based human-AI conversations.” *ICED conference*. in submission.