How Generative AI Image Generation Aligns with and Shapes Human Mental Images

Gwendolyn Ghittini
University of St. Gallen
St. Gallen, Switzerland
gwen.ghittini@student.unisg.ch

Leon Reicherts
University of St. Gallen
St. Gallen, Switzerland
leon.reicherts@unisg.ch

Although Generative AI (GenAI) image generators are widely used by people to turn an idea into a picture within seconds, the process of generating and iterating over these images based on people's mental images remains underexplored. We report an exploratory mixed-methods study combining a pre-survey of everyday GenAI image-generation practices ($N = 156$) with an image-generation task ($N = 46$) in which participants generated images of an assigned abstract or concrete target word and repeatedly compared the output with their mental image. Participants rather quickly converged on an AI-generated image that matched their mental image, with further iterations refining rather than reshaping it, but this convergence played out differently depending on what was being visualised: abstract concepts left more room for the AI to shape the outcome, while concrete concepts anchored participants to a more stable image that resisted further AI influence. We discuss implications for designing GenAI image-generation interfaces around this distinction to better support the different processes of expressing and updating one's internal images.

Generative AI, text-to-image generation, mental imagery, iterative prompting, human-AI co-creation, AI authorship

1. INTRODUCTION

Externalising a mental image has traditionally required slow, skill-dependent processes such as sketching, describing, or building physical mock-ups. GenAI image generators change this by letting users produce visual outputs through natural-language prompts rather than manual drawing (Oppenlaender 2022; Liu and Chilton 2022). Tools like ChatGPT that also offer such image generation capabilities are used for a plethora of activities across private, educational, and work contexts (Wolf and Maier 2024; Hannig et al. 2026). Yet, how ordinary users draw on GenAI to visualise their own mental images, and what happens to those mental images once an AI-generated version of them exists, remains largely unstudied.

Prior HCI research has examined how GenAI may support visual artists' ideation and communication (Ko et al. 2023), community-based collaborative speculation (Epstein et al. 2022), and team-based creative design (Han et al. 2024). In the present study, we instead examine how the image-generation process interacts with people's mental images depending on whether concrete or abstract concepts are represented.

This paper reports a mixed-methods study combining a pre-survey on everyday GenAI image-generation practices with an image-generation task consisting of three iterations in which participants compared AI outputs to their mental image on every round. We focus on the results with the clearest HCI implications, organised around three research questions: how agreement between AI output and mental image develops across iterations (RQ1); how prompting strategies and perceived mental-image change differ between iterations (RQ2); and how these patterns diverge between *abstract* and *concrete* target concepts (RQ3).

2. RELATED WORK

Mental imagery is commonly defined as "seeing in the absence of the appropriate immediate sensory input" and can arise either from briefly retained sensory input or from information activated in long-term memory (Kosslyn et al. 1995). Individual differences in imagery vividness are commonly assessed with the Vividness of Visual Imagery Questionnaire (VVIQ; Marks 1973); most people report broadly typical imagery ability, while small minorities report absent or weak imagery (aphantasia or hypophantasia) or unusually vivid imagery (hyperphantasia) (Wright et al. 2024).

© Ghittini and Reicherts. Published by BCS Learning and Development Ltd.

External representations (e.g. drawings and diagrams) can form part of the cognitive process rather than merely recording it, enabling people to inspect, reorganise, and develop ideas (Kirsh 2010; Scaife and Rogers 1996). GenAI image generators are a novel case of such an external representation: unlike sketching, they let a mental image be externalised almost instantly and repeatedly, through prompting rather than manual rendering.

Prompt formulation is treated as a design practice in its own right (Oppenlaender 2022, 2024). In many text-to-image systems, text prompts are the primary means through which users express creative intent, but translating that intent into a satisfactory output often requires iterative trial and error (Liu and Chilton 2022; Brade et al. 2023). Across LLM-based chatbot design and text-to-image generation, non-experts and beginners can struggle to develop effective prompts: users may explore opportunistically rather than systematically and have difficulty discovering useful keywords (Zamfirescu-Pereira et al. 2023; Brade et al. 2023). Emotional or contextual nuance can be especially challenging to convey (Han et al. 2024). Prompting is therefore often iterative: users add, delete, or replace terms across successive prompts within a usage session (Xie et al. 2023).

Another strand of research examines how users experience and relate to GenAI outputs. In one qualitative study, responses to GenAI images extended beyond appraisal of the output to how participants positioned themselves in relation to the system. When confronted with unsettling or unfamiliar outputs, participants adopted varied "relational strategies", positioning themselves as superior or inferior to GenAI or revising their own interpretations in response to it (Rapp et al. 2025). Collaborative studies similarly find that users treat GenAI as a useful but subordinate tool, valuable for labour-intensive or exploratory tasks, while remaining wary of it displacing their own ideas (Han et al. 2024). This ambivalence extends to authorship: because the final image reflects the user's prompt, the model's training data, and the model's own interpretation, authorship is better understood as relational and distributed across human and system rather than belonging to either alone (Bomba and De Angeli 2025). Other work has framed GenAI as a catalyst for combinatorial innovation, recombining concepts and ideas while acting as a complementary artefact that augments human creativity and innovation (Krakowski 2025).

GenAI image generation is thus iterative, prompt-mediated, and perceived as co-created. Exposure to AI-generated imagery can shape subsequent visual ideation (Wadinambiarachchi et al. 2024) and other systems have used generated images as feedback in iterative mental-image retrieval (Yang et al. 2025). Whether and how people's *own* mental image is affected by repeatedly comparing it against GenAI output – and whether this differs for more vs. less visually concrete ideas – remains open; the present study addresses this gap.

3. METHOD

The study used an exploratory mixed-methods design consisting of a pre-survey and a three-round image-generation task, both administered online via Qualtrics. The study received ethics approval from the authors' university, and all participants provided informed consent. Participants were recruited through convenience sampling, mainly via the lead author's social network. The participant sample consisted of healthy adults located in Europe. The pre-survey received $N = 189$ valid responses, of which $N = 156$ (82.5%) had at least some prior GenAI image-generation experience, which was required for the main study; all pre-survey results reported below refer to this subsample. The pre-survey captured everyday visualisation habits, prompting practices, and perceptions of GenAI image tools. Participants who reported prior image-generation experience and volunteered an email address were invited to the main study; $N = 46$ completed it (20 female, 26 male; $M_{\text{age}} = 28.3$, $SD = 7.59$).

In the main study, participants first completed the 16-item VVIQ (range 16–80, higher scores indicating more vivid imagery; $M = 59.1$, $SD = 8.6$). The sample mean closely matched the mean of 58.6 ($SD = 13.3$) reported in a community sample by Zeman et al. (2020). VVIQ categories distinguish aphantasia (score 16), hypophantasia (17–32), typical imagery ability (33–74), and hyperphantasia (75–80) (Wright et al. 2024). Participants then received one randomly assigned target word from a set of 19 abstract and concrete nouns adapted from Rapp et al. (2025), each word assigned to 2–4 participants (e.g. *justice*, *nostalgia* vs. *glacier*, *mirror*; $n = 23$ cases per category; full list in Appendix B). For each of three iterations, participants were asked to construct a mental image of the word, formulated a prompt expressing it, and used an image-generation model of their choice to generate a corresponding image (Figure 1). Allowing participants to use a familiar model reduced the need to learn a new tool and reflected our focus on prompting behaviour across iterations rather than comparisons between image outputs. After each iteration, participants rated (i) overall agreement between the AI image and their mental image, (ii) agreement on three subscales (composition, atmosphere/style, detail), and (iii) how much their mental image had changed

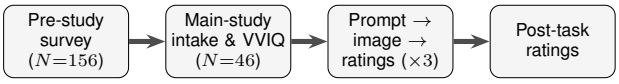

Figure 1: *Study procedure: pre-survey on personal AI use and experience ($N = 156$), followed by the main study ($N = 46$) consisting of three rounds of prompting/generation/rating, and final source, authorship, and preference ratings.*

Table 1: *Prompt-coding dimensions used to analyse the 138 prompts (46 cases × 3 rounds).*

Dimension	Categories
Function	F1 generation; F2 local edit; F3 full re-generation; F4 meta/repair
Control	C1 low control (basic motif); C2 medium (setting/perspective/style); C3 high (precise spatial cues, style, constraints)
Vibe	mood/atmosphere terms (e.g. warm, eerie) vs. concrete visual elements only

since the previous round, all on 7-point scales. After the third iteration, participants rated the perceived source (own imagination vs. AI), the authorship split, and their preference for the AI image vs. their own mental image.

Participants' prompts were additionally coded independently by two researchers using a deductively developed scheme (Table 1): a *function* code per round (generation, edit, re-generation, or repair), a three-level *prompt-engineering* rating capturing how precisely subject, composition, and style were specified, and a binary *"vibe"* code marking prompts that specified mood or atmosphere rather than only concrete visual elements (Table 1 gives examples). Initial percentage agreement between the two coders was 93%; remaining disagreements were discussed until consensus was reached, in some cases consulting ChatGPT as an auxiliary third reference point (two-out-of-three logic) where needed, with final decisions always remaining under human control. Round-level function codes were further grouped into five ordered three-round sequence types (e.g. *iterative refinement*, F1→F2→F2).

Since most outcome variables are ordinal self-report scales, we mainly report medians (Md), interquartile ranges (IQR), modal categories, and full response distributions rather than means. Some exceptions are participant age, the validated VVIQ instrument, and the two 0–100 percentage-allocation sliders for perceived AI contribution and AI authorship, which we report as mean and SD; the analysis throughout is descriptive/exploratory rather than inferential.

4. RESULTS

The results address the research questions on iteration-level agreement (RQ1, Section 4.3), prompting and perceived change (RQ2, Sections 4.2 and 4.4), and abstract/concrete divergence (RQ3) throughout. The first subsection reports on the pre-study subsample ($N = 156$; $M_{\text{age}} = 30.5$, $SD = 8.4$; 70 female, 86 male); the remaining subsections report on the main study ($N = 46$; $M_{\text{age}} = 28.3$, $SD = 7.6$; 20 female, 26 male).

4.1. Everyday use of AI image generation

All 156 participants in the analysed pre-study subsample had prior experience with GenAI image generation; 63.5% currently used such tools at least monthly, while 36.5% used them less frequently. Participants reported needing few iterations for a satisfying result in daily use (2–3: 48.1%; 4–6: 35.9%), which motivated the three-iteration design of the main task. Most participants (84%) described GenAI image generators as a *tool*, with far fewer selecting *creative partner* (25%) or *co-author* (7.7%) on this multi-select item – foreshadowing the authorship results further below. Consistent with this instrumental framing, 75.6% agreed that GenAI inspires ideas they would not have had otherwise ($Md = 6$, IQR 5–6; on a 7-point scale), yet only 45.5% agreed that the AI understands their intentions ($Md = 4$, IQR 3–5). Prompt phrasing itself was rated as moderately difficult ($Md = 4$, IQR 3–5; on a 7-point scale). Within the subsample who completed the main study ($N = 46$), the same item received equivalent ratings ($Md = 4$, IQR 3–5; Table 2). The most commonly reported challenges were unexpected or incorrect interpretations by the model (61.5%), controlling details (57.1%), and finding the right words (53.2%). The most frequently reported concerns were privacy/data handling (19.9% of open responses) and realism-enabled deception such as deepfakes (17.3%). Control over the output and alignment with the mental image were both rated among the highest priorities ($Md = 6$ for both; IQR 5.75–7 and 5–7, respectively; on a 7-point scale).

4.2. Prompting is mostly incremental refinement

Across the 138 coded prompts, participants mainly used iterative refinement (function sequence F1→F2→F2, 73.9% of cases) rather than repeatedly re-specifying the image from scratch (F1→F3→F3, 15.2%); this pattern held broadly for both categories, though abstract-word participants were notably more likely to include at least one full re-specification round in their sequence (30.4% vs. 17.4% of concrete-word participants; Figure 3). Prompt-engineering effort was moderate: 47.1% of all

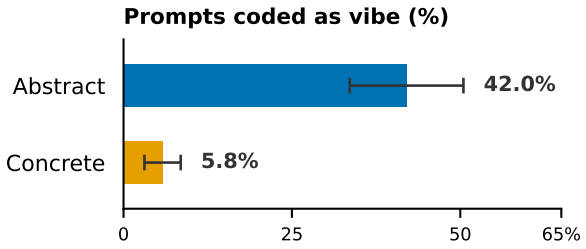

Figure 2: *Percentage of prompts coded as "vibe" prompts (mood/atmosphere language) by target-word category, pooled across the three iterations. Error bars show ±1 SE computed across participants' per-participant vibe proportions ($n = 23$ per category).*

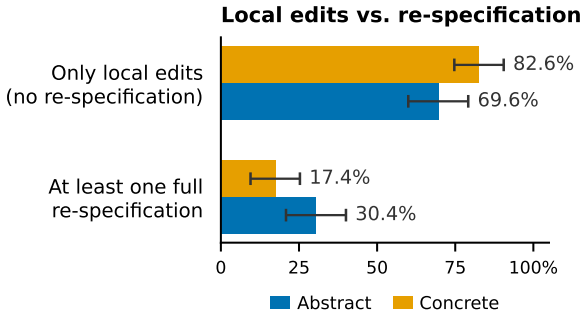

Figure 3: *Percentage of participants whose three-round prompt sequence included at least one full re-specification round (function code F3) vs. only local edits. Error bars show ±1 SE (binomial), $n = 23$ per category.*

prompts were coded C1 (low control) and only 23.2% C3 (high control). The clearest divide, however, was in *how* participants described their target: 42.0% of prompts for abstract words were coded as "vibe" prompts (specifying mood or atmosphere) compared with only 5.8% for concrete words (Figure 2). Abstract concepts, lacking a stable visual referent, were translated into subjective, atmospheric language, whereas concrete concepts were mainly described by their visible features.

A similar divide was already present in the very first prompt: only 21.7% of round-1 abstract-word prompts were coded C1 (low control), vs. 39.1% of round-1 concrete-word prompts, meaning participants more often needed to go beyond a bare, minimal description just to convey an abstract concept in the first place. Sustained high-effort prompting showed a different pattern: no abstract-word participant used the highest-control code (C3) in all three rounds, whereas 13.0% of concrete-word participants did. These patterns show that abstract concepts elicited greater initial elaboration, whereas only concrete concepts led some participants to sustain highly controlled prompting across all three rounds.

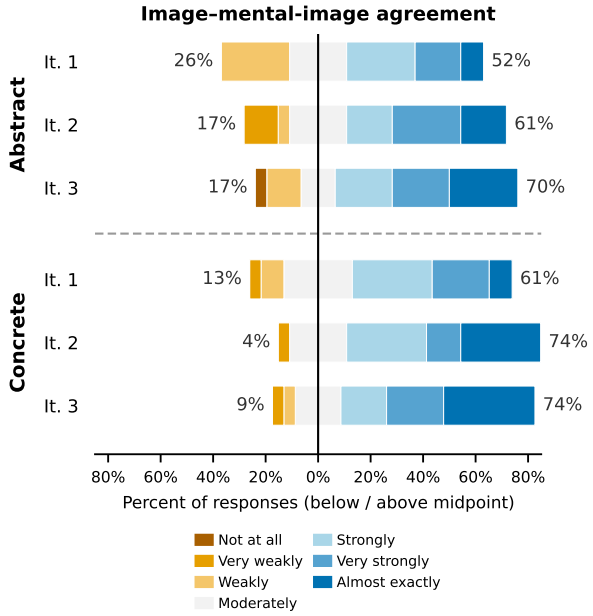

Figure 4: *Likert plot of overall image–mental-image agreement (7-point scale), across the three iterations, by target-word category.*

4.3. Agreement between mental and AI images

Overall agreement between the AI-generated image and participants' mental image was already moderate-to-high in the first iteration and increased only modestly thereafter ($Md = 5$, IQR 4–6, in iterations 1–2, rising to $Md = 6$, IQR 4–7, in iteration 3). Figure 4 shows this as a Likert plot per category, with the percentage of responses below the scale midpoint ("Moderately", 4) plotted left of the zero line and the percentage above it to the right. For concrete words, the share of responses above the midpoint was already 61% in iteration 1 and reached 74% by iterations 2–3. Abstract words started lower (52% above the midpoint in iteration 1) but "caught up" by iteration 3 (70%). Agreement for abstract words rose continuously, whereas agreement for concrete words increased between iterations 1 and 2 and then plateaued; by iteration 3, the two categories were similar.

The other items on agreement with specific aspects – *atmosphere/style* ($Md = 6$, IQR 5.25–6.75) and *important-detail agreement* ($Md = 6$, IQR 5–6) – both remained at $Md = 6$ from iteration 1 onward, for abstract and concrete words alike. *Compositional agreement* improved slightly across iterations for abstract words ($Md = 5$, IQR 4–6, to $Md = 6$, IQR 5–7), whereas concrete words were already at $Md = 6$ (IQR 4–6) from iteration 1 onward.

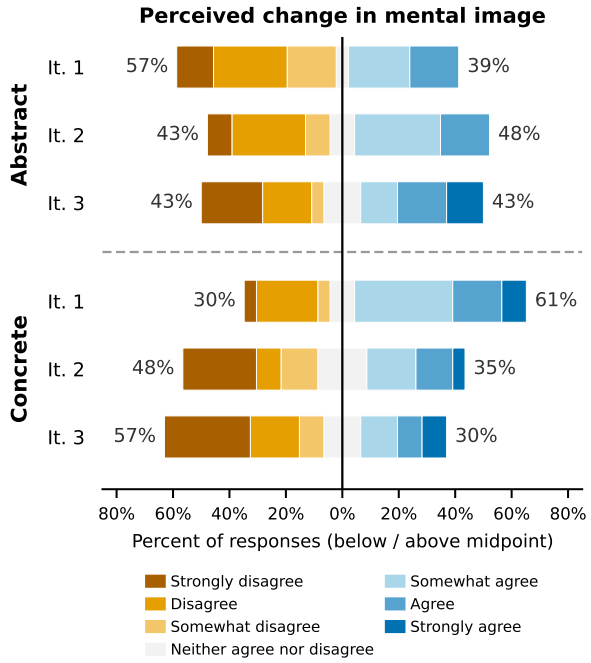

Figure 5: *Likert plot of perceived change in the mental image (7-point agreement scale), across the three iterations, by target-word category.*

4.4. Perceived change in mental images

Although the AI image increasingly matched what participants reported imagining, this did not necessarily mean that their mental images had stopped changing. Perceived change declined overall across iterations ($Md = 4.5$, IQR 2–5, in iteration 1 to $Md = 3.5$, IQR 1.25–5, in iteration 3), consistent with a "stabilisation process" in which the first AI output had the largest influence and later iterations mainly refined it. However, Figure 5 shows that this decline is not uniform across target-word categories. For concrete words, agreement with the statement "my mental image changed" fell from 61% (above the midpoint) in iteration 1 to 30% by iteration 3 – a clear, monotonic "stabilisation". For abstract words, the pattern was non-linear: 39% above the midpoint in iteration 1, rising to 48% in iteration 2, before settling at 43% in iteration 3. Concrete concepts thus appear to provide a firmer internal reference point that seems to quickly resist further AI influence, while abstract mental images remain comparatively open to revision even after two rounds of feedback.

4.5. Source, authorship, and type of change

After the final round, participants reported the source of their current mental image, rated AI contribution and authorship on two 0–100 sliders, and selected the type(s) of change that happened most often. Source attribution differed by category. For abstract

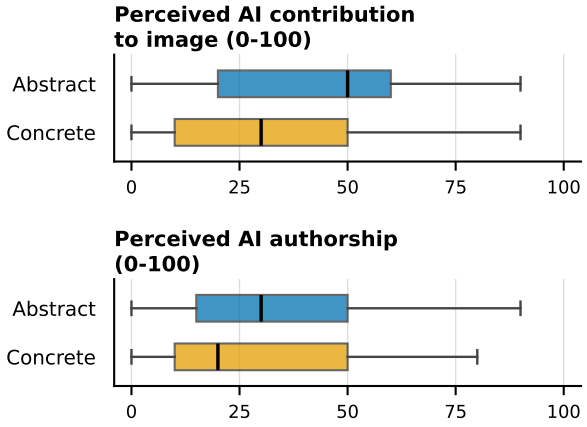

Figure 6: *Distributions of perceived AI contribution to the final mental image and perceived AI authorship, by target-word category.*

words, 60.9% described their mental image as a blend, compared with 30.4% as mostly their original imagination and 8.7% as mostly AI-based. For concrete words, mostly original imagination was the modal response (47.8%), followed by blend (30.4%) and mostly AI-based (21.7%). The slider responses differed only modestly: mean AI contribution was 39.1% for abstract and 34.8% for concrete words, while mean AI authorship was 35.2% and 30.9%, respectively (Figure 6 shows the corresponding distributions, with centre lines indicating medians). Across both categories, participants attributed 33.0% of authorship to the AI on average ($SD = 24.1$), and thus continued to see themselves as the primary authors of their mental images. The reported type of change reinforced this category distinction. Although participants in both categories selected "small tweaks to details" more often than "replacing major parts of my image with the AI's version," abstract-word participants selected small tweaks considerably more often (78.3% vs. 47.8%), whereas concrete-word participants selected major replacement twice as often (17.4% vs. 8.7%). Together with the source responses, this suggests that AI input was more often incorporated into abstract mental images through incremental blending, while concrete mental images were more often preserved as primarily the participant's own but were also somewhat more likely to undergo major replacement (see Figure 7 in the Appendix).

4.6. Preference for own or AI-generated images

Despite three rounds of AI-assisted image generation, participants generally preferred their own mental image to the AI-generated output, with the same $Md = 3$ for both categories (on the 1–7 scale, where 1 = strongly prefer own image, 4 = no preference, and 7 = strongly prefer AI image).

Concrete-word participants' responses were more dispersed than abstract-word participants' responses (IQR 1–5 vs. 2–5). Moreover, a larger share of concrete-word participants preferred their own mental image (65.2% vs. 52.2%). This more dispersed pattern is consistent with the source-attribution split, whereas abstract-word responses included more neutral preferences (see Figure 7).

5. DISCUSSION

Our study shows that GenAI image generation functions like an iterative feedback loop: participants externalised an initial mental image, evaluated the AI's visual interpretation, and selectively incorporated or discarded elements of it into their own mental image, largely through small, cumulative edits rather than wholesale replacement. This supports treating GenAI-generated images as external representations in the sense of Kirsh (2010) and Scaife and Rogers (1996): artefacts that do not just record a thought but actively participate in refining it.

The strongest HCI implication is that how much room the AI had to shape the outcome depended on what participants were visualising. Abstract-word participants more often used mood-based "vibe" prompts and were more likely to fully re-specify their prompt rather than make only local edits (30.4% vs. 17.4%). They also attributed a larger share of the result to the AI on the contribution and authorship sliders. The iteration trajectories reinforce this contrast. For concrete words, reported mental-image change *fell* from 61% in iteration 1 to 30% in iteration 3, while above-midpoint agreement *rose* from 61% to 74%, indicating faster agreement and stabilisation of the mental image. For abstract words, agreement also *rose* (from 52% to 70%) while mental-image change had a slight tendency to *rise* (at least from the first to the second iteration), indicating that despite increasing agreement, the mental image remained more "malleable". These changes over iterations may therefore signal more quickly stabilising concrete images versus ongoing adaptation around abstract ones. This interpretation aligns with prior work that frames prompt engineering as search through a design space (Liu and Chilton 2022) and documents exploratory changes across prompt sessions (Xie et al. 2023). In our data, such exploratory adaptation appears particularly pronounced for abstract images, although participants in both conditions reported some mental-image change.

These differences suggest that interfaces could adapt to these two modes of use. For concrete referents, they could offer region- or attribute-level editing tools that let users adjust a specific object, colour, or spatial detail and "lock" a direction that already feels right. For abstract or emotional content, interfaces might foreground the atmospheric, mood-based vocabulary users already reach for, for example through mood or tone controls, while retaining exploratory alternatives rather than assuming the user is ready to settle on one image – as these were the aspects found to be prevalent in the prompts for abstract vs. concrete referents. Because atmosphere/style agreement was already high in the first round, whereas compositional agreement started lower and improved across rounds – particularly for abstract words – interfaces could prioritise controls for layout and spatial relations.

Limitations and Future Work. The main-study sample was relatively small ($N = 46$, 23 per category); all key measures were self-reported, and the generated images were not evaluated. Accordingly, our descriptive results on how mental images *changed* need to be understood as reports of *perceived change* rather than evidence of objectively verified representational change. Future work should replicate the comparison with larger, more diverse samples, validate prompt taxonomies, and pair self-reported mental-image change with independent measures (e.g. sketch-based tasks) to test whether the reported stabilisation reflects genuine representational change rather than response-scale effects.

6. CONCLUSION

Returning to our three research questions: across three rounds of iterative prompting, agreement between AI output and mental image increased only modestly beyond the first iteration (RQ1), while participants mainly refined their prompts incrementally and reported small adjustments to their mental images rather than wholesale replacement (RQ2). The clearest pattern, however, was not the iteration effect itself but its interaction with what participants were visualising (RQ3): abstract, less visually fixed concepts invited mood-based prompting and a larger role for the AI, while concrete concepts anchored participants to a stable mental image that more quickly resisted further AI influence. Designing GenAI image-generation interfaces around this distinction – rather than one generic prompting paradigm – is an opportunity for future HCI work on human-AI co-creation.

7. ACKNOWLEDGMENTS

We thank Johannes Schöning for his valuable input at key points during the project, all participants for their time and contributions, and the reviewers for their constructive feedback.

REFERENCES

Bomba, F. and De Angeli, A. (2025) Agency and Authorship in AI Art: Transformational Practices for Epistemic Troubles. *International Journal of Human-Computer Studies*, 205, 103652.

Brade, S., Wang, B., Sousa, M., Oore, S. and Grossman, T. (2023) Promptify: Text-to-Image Generation through Interactive Prompt Exploration with Large Language Models. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–14.

Epstein, Z., Schroeder, H. and Newman, D. (2022) When Happy Accidents Spark Creativity: Bringing Collaborative Speculation to Life with Generative AI. *Proceedings of the 13th International Conference on Computational Creativity*, 334–338.

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

Hannig, L., Bush, A., Aksoy, M., Trappen, T., Becker, S. and Ontrup, G. (2026) Campus AI vs. Commercial AI: Comparing How Students and Employees Perceive Their University's LLM Chatbot vs. ChatGPT. *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–23.

Kirsh, D. (2010) Thinking with External Representations. *AI & Society*, 25(4), 441–454.

Ko, H.-K., Park, G., Jeon, H., Jo, J., Kim, J. and Seo, J. (2023) Large-Scale Text-to-Image Generation Models for Visual Artists' Creative Works. *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 919–933.

Kosslyn, S. M., Behrmann, M. and Jeannerod, M. (1995) The Cognitive Neuroscience of Mental Imagery. *Neuropsychologia*, 33(11), 1335–1344.

Krakowski, S. (2025) Human-AI Agency in the Age of Generative AI. *Information and Organization*, 35(1), 100560.

Liu, V. and Chilton, L. B. (2022) Design Guidelines for Prompt Engineering Text-to-Image Generative Models. *CHI Conference on Human Factors in Computing Systems*, 1–23.

Marks, D. F. (1973) Visual Imagery Differences in the Recall of Pictures. *British Journal of Psychology*, 64(1), 17–24.

Oppenlaender, J. (2022) The Creativity of Text-to-Image Generation. *Proceedings of the 25th International Academic Mindtrek Conference*, 192–202.

Oppenlaender, J. (2024) A Taxonomy of Prompt Modifiers for Text-to-Image Generation. *Behaviour & Information Technology*, 43(15), 3763–3776.

Rapp, A., Di Lodovico, C., Torrielli, F. and Di Caro, L. (2025) How Do People Experience the Images Created by Generative Artificial Intelligence? An Exploration of People's Perceptions, Appraisals, and Emotions Related to a Gen-AI Text-to-Image Model and Its Creations. *International Journal of Human-Computer Studies*, 193, 103375.

Scaife, M. and Rogers, Y. (1996) External Cognition: How Do Graphical Representations Work? *International Journal of Human-Computer Studies*, 45(2), 185–213.

Wadinambiarachchi, S., Kelly, R. M., Pareek, S., Zhou, Q. and Velloso, E. (2024) The Effects of Generative AI on Design Fixation and Divergent Thinking. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–18.

Wolf, V. and Maier, C. (2024) ChatGPT Usage in Everyday Life: A Motivation-Theoretic Mixed-Methods Study. *International Journal of Information Management*, 79, 102821.

Wright, D. J., Scott, M. W., Kraeutner, S. N., Barhoun, P., Bertollo, M., Campbell, M. J., Waltzing, B. M., Dahm, S. F., Esselaar, M., Frank, C., Hardwick, R. M., Fuelscher, I., Marshall, B., Hodges, N. J., Hyde, C. and Holmes, P. S. (2024) An International Estimate of the Prevalence of Differing Visual Imagery Abilities. *Frontiers in Psychology*, 15, 1454107.

Xie, Y., Pan, Z., Ma, J., Jie, L. and Mei, Q. (2023) A Prompt Log Analysis of Text-to-Image Generation Systems. *Proceedings of the ACM Web Conference 2023*, 3892–3902.

Yang, D., Liu, M., Lo, C.-H., Zhang, Y. and Davis, J. (2025) GenIR: Generative Visual Feedback for Mental Image Retrieval. *Advances in Neural Information Processing Systems*, 38 (Main Conference), 137210–137239.

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B. and Yang, Q. (2023) Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–21.

Zeman, A., Milton, F., Della Sala, S., Dewar, M., Frayling, T. M., Gaddum, J., Hattersley, A. T., Heuerman-Williamson, B., Jones, K., MacKisack, M. and Winlove, C. (2020) Phantasia–The Psychological Significance of Lifelong Visual Imagery Vividness Extremes. *Cortex*, 130, 426–440.

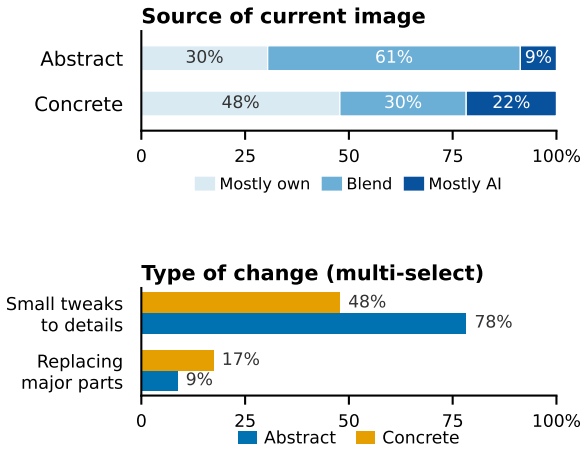

Figure 7: *Source of the current image (single-choice, all options), and two type-of-change options (multi-select) with the largest abstract/concrete contrast, by target-word category.*

A. AI USE DECLARATION

We used ChatGPT and Claude under close guidance and oversight to rephrase and summarise this paper's underlying project report and to improve the writing. These tools were not employed to generate original ideas or data. We further used spell-checking tools like Overleaf's spell checker and Grammarly.

B. PARTICIPANT DEMOGRAPHICS AND AI USE

Table 2 gives an overview of main-study participants' demographics (age, gender, and VVIQ) and prior GenAI image-generation experience. All of them had previously used a GenAI image generator, consistent with the recruitment criterion described in Section 3.

C. TARGET WORDS

Table 3 lists the 19 target words assigned to main-study participants ($N = 46$; $n = 23$ per category), with the number of participants who received each word.

D. EXAMPLE PROMPT SEQUENCES ACROSS ITERATIONS

To illustrate how prompts evolved across the three rounds, we reproduce verbatim prompt sequences from four participants, two per target-word category. Function (F1–F4) and control (C1–C3) codes follow the coding scheme in Table 1. Prompts appear as recorded.

P1 (concrete, *House*). *R1 (F1, C1):* "Create a basic 2D drawing of a house only consisting of black straight lines. In addition to the walls and roof, it shows a door and a chimney." *R2 (F3, C2):* "Create a basic 2D drawing of a house only consisting of black straight lines – no double lines or any details. In addition to the walls and roof, it shows a small door on the bottom left corner and a chimney on top right." *R3 (F3, C3):* "Create a basic 2D drawing of a house only consisting of black straight lines – no double lines or any details. Walls and roof form one line – the roof does not exceed the walls. In addition to the walls and roof, it shows a small door with a black knob on the bottom left corner and a chimney on top right. The chimney only consists of three lines – no double lines or extras. The house is narrow compared to its width – like a 3 story building."

P2 (concrete, *Steering Wheel*). *R1 (F1, C1):* "Draw me a steering wheel which is black and from the brand opel. Just the steering wheel and no surroundings. It is from my perspective as a driver" *R2 (F2, C2):* "Please add more details so that it looks like i am sitting in the car and holding the steering wheel with both hands. So it's not just the steering wheel but i sit in the car, hold it and also add the other details which are usually in the car. But focus is on the steering wheel and not what is outside" *R3 (F2, C2):* "Leave the inside exactly like this but add an atmosphere outside of the car. No details just a slightly atmosphere. It is summer outside, a bit green and i am driving in the nature"

Table 2: *Main-study participant demographics and prior GenAI image-generation experience ($N = 46$).*

Variable	Value
Age (years), M (SD)	28.3 (7.59); range 20–66
Gender	26 male, 20 female
Target-word category	23 abstract, 23 concrete
VVIQ score, M (SD)	59.1 (8.6)
Prior duration of GenAI use	7–12 mo.: 34.8% 1–2 yr: 32.6% $>$2 yr: 17.4% 1–6 mo.: 13.0% $<$1 mo.: 2.2%
Current frequency of GenAI use	Less than monthly: 41.3% Weekly: 26.1% Monthly: 15.2% Daily: 10.9% Several times/week: 2.2% No longer use it: 4.3%
Prompting difficulty (7-point scale)	$Md = 4$ (IQR 3–5)

Table 3: *Target words used in the main study, by category, with the number of participants assigned each word.*

Abstract (n)	Concrete (n)
Anarchy (2)	Glacier (3)
Anger (2)	House (3)
Euphoria (3)	Lake (2)
Infinity (2)	Man (2)
Justice (4)	Margarita (3)
Laziness (2)	Microscope (2)
Love (2)	Mirror (3)
Luck (2)	Steering Wheel (2)
Nostalgia (2)	Wig (3)
Wisdom (2)	

P3 (abstract, *Anger*). *R1 (F1, C1):* "I need you to generate my mental image of Anger which is a blazing fire in multiple colors (red, orange and yellow). It is surronded by stones to keep it contained. Otherwise, there is not much around the fire" *R2 (F2, C2):* "The fire needs to look more hostile, more like there could be danger with this fire. It looks more like a campfire at the moment and it needs to look more dangerous" *R3 (F2, C2):* "The fire needs to look more dangerous. the image needs to represent a fire that is no longer safe to be around and that the people who are in the immediate vicinity or in the surrounding area need to evacuate right away so that the fire doesn't get to them"

P4 (abstract, *Euphoria*). *R1 (F1, C3):* "Cinematic, photorealistic scene of a young woman standing alone in a softly glowing environment, surrounded by neon purples, deep blues, and subtle pink highlights. The atmosphere is slightly hazy, with tiny sparkling particles floating in the air, catching the light. Everything feels slowed down and weightless, as if time is suspended. Her expression is calm yet emotionally intense, fully immersed in the moment. Soft lighting wraps around her, creating a dreamy, surreal mood with high detail, shallow depth of field, and an 85mm lens aesthetic." *R2 (F2, C1):* "make it more euphoric feeling, like she is floating in a galaxy and the girl looks ethereal" *R3 (F2, C1):* "looks good but its not giving me a fully euphoric feeling, make it embody an out of body experience"

The concrete-word sequences show contrasting refinement styles: P1 tightens an already precise motif through stricter spatial constraints, whereas P2 adds surrounding context (driver, car interior, outdoor atmosphere) across rounds. Both remain feature-based throughout. P3 shifts into mood-driven "vibe" language across rounds (Section 4.2); P4 uses it from the first round and later relies on short, evocative cues rather than additional visual detail.