VariationWeaver: Scaffolding Designerly Exploration and Convergence with a Text-to-Image Model

ANONYMOUS AUTHOR(S)
SUBMISSION ID: 8966

Fig. 1. *VariationWeaver*: An AI-enabled product design interface for novices. The components include (A) Prompt Box, (B) Dimension Palette, (C) Image Canvas, and (D) Open/Close Side Bar, Add Comment, Favorite Folder Buttons

ABSTRACT

Generative image models have empowered people to visually represent new product ideas. However, from a design perspective, the speed and fidelity of image models could induce premature fixation on particular solutions or lead to a shallow understanding of a design space, especially for novices who have not internalized effective practices. Our research investigates how to foreground design-space dimensionality and provide more support for divergent and convergent thinking when interacting with an image model. We present *VariationWeaver*, a custom prototype for a text-to-image interface for novice designers that surfaces and explains design terminology, renders the fidelity of output images relative to prompt specificity, and supports variation and comparison along key design variables. An exploratory study with 15 novice designers characterizes how terminology scaffolds support dimensional acclimation, how variation with aligned comparison and staged fidelity shape exploration and choice, and how participants experienced confidence and authorship.

CCS Concepts: • **Human-centered computing** → **Empirical studies in interaction design**; **Text input**; *Graphical user interfaces*; *Collaborative interaction*; Natural language interfaces.

Additional Key Words and Phrases: Creativity support tools, design ideation, variation, human-AI interaction, text-to-image models

ACM Reference Format:

Anonymous Author(s). 2026. VariationWeaver: Scaffolding Designerly Exploration and Convergence with a Text-to-Image Model. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 30 pages. https://doi.org/10.1145/XXXXXXX.XXXXXXX

1 INTRODUCTION

Modern generative image models can produce photorealistic concepts from a single prompt. While empowering, high-fidelity renderings can induce fixation [34], draw attention to surface features rather than underlying dimensions [11], and hinder both broad divergence and principled convergence. In groups, polished images may signal premature closure and dampen critique [11, 20, 21]. Difficulties compound in *asynchronous* work, where progress depends on artifact-mediated handoffs and rationale often dissipates across tools, places, and time [27, 52, 64, 85].

We ground our approach in three strands. *Information Foraging* suggests people follow cues with strong "information scent," so interfaces that surface high-yield cues can help novices learn unfamiliar spaces [61]; recent systems show how exposing domain language and principles supports this process [54, 55]. *Prototyping research* clarifies fidelity trade-offs: sketchy artifacts invite exploration and critique, whereas higher fidelity supports specification and judgment; tuning fidelity by stage mitigates premature commitment [11, 67, 80]. *Variation Theory* argues that people discern critical features via systematic contrast—varying one factor while holding others fixed [24, 49]; aligned, like-with-like comparison makes correspondences legible [25] and supports explicit trade-offs [60]. In practice, parallel prototyping improves critique quality and downstream decisions [20, 21, 77].

Recent HCI systems broaden interaction beyond chat by scaffolding *expression* (prompt templates, semantic labels, transparent pipelines) [37, 81, 82] and by supporting *iteration* through layered/region editing with before–after previews [17, 66, 70]. On the output side, interfaces scale browsing with parameter sweeps [48], clustering/spatial layouts [7], and dimension-aware exploration [74]; shared, artifact-anchored views can help build common ground [28, 84].

Despite these advances, three gaps persist. First, the design space often remains *tacit*: alternatives are generated without naming the underlying dimensions or terms novices need [41]. Second, comparisons are not *commensurate* across iterations, obscuring trade-offs [25, 60]. Third, rationales rarely *travel* with artifacts, limiting asynchronous progress [64, 85].

We introduce *VariationWeaver*, an interactive canvas for text-to-image product ideation that (a) surfaces in-place terminology to make dimensions explicit, (b) supports controlled one-factor variation with aligned side-by-side comparison [25, 49], and (c) stages fidelity from sketch-like to detail-preserving as specificity increases [11, 67, 80]. The canvas aims to produce commensurate, artifact-anchored views so alternatives are comparable and rationale can move across handoffs [28, 60, 84].

We conducted a qualitative study with 15 novices to investigate:

RQ1: Dimensional acclimation: How did novices use embedded terminology to acclimate to the design space?

RQ2: Variation, comparison, and fidelity: How did novices use variation and aligned comparison—together with staged fidelity—to explore alternatives and make deliberate design choices?

RQ3: Experience: How did novices feel about using *VariationWeaver* for product design?

Across sessions, three patterns recurred: (1) *visual language scaffolding* helped novices name dimensions and steer exploration; (2) *controlled variation* shown in aligned, side-by-side views made differences—and thus trade-offs—explicit;

and (3) *sketch-first rendering* supported ideation but hampered evaluation until fidelity rose with intent. Confidence and ownership increased when edits took effect in the images and when a colleague's artifacts provided a contrasting reference point.

This paper contributes:

- **System:** *VariationWeaver*—a working canvas that operationalizes three designerly principles for T2I ideation: make dimensions and terminology visible in place; couple controlled, one-factor variation with aligned comparison; and stage output fidelity to match evolving intent.
- **Empirical insights:** A nuanced account of *when and why* these principles help or hinder novices—showing conditions that enable dimensional acclimation, deliberate trade-offs, and felt authorship—and distilling design guidance for future GenAI ideation tools.

2 RELATED WORK

We position our work across four areas: (1) the power and pitfalls of modern generative pipelines; (2) HCI responses for expression, iteration, and scale; (3) designerly lenses—fidelity, dimensionality, and structured variation; and (4) comparison and collaborative sensemaking for convergence and handoffs.

2.1 Generative Image Models: Power and Pitfalls

Instruction-tuned LLMs produce structured text briefs [10, 16, 53]. Diffusion pipelines support image generation and editing [12, 63, 65]; related work extends to 3D assets [42, 44, 62] and to video and interactive media [3, 4, 30, 38, 71]. Access to high-fidelity visuals can narrow exploration via fixation and surface-level attention [11, 34]; early ideation benefits from sketching and structured cues [51, 83]. In collaboration, polished images may signal finality and dampen critique [20, 21]. Commensurate representations and portable rationale support principled judgment and handoff [60, 85]. Text-to-image artifacts can also catalyze shared semantics without collapsing options [43], though strategy-centric guidance may homogenize ideas and weaken later originality [39]. Preserving intent across roles during convergence remains critical [58].

2.2 HCI Responses: Rethinking Interaction With Models

For **expression**, systems scaffold prompting with templates [82], semantic labels [37], and transparent stepwise pipelines for control/debugging [81]; prompt-engineering aids further support refinement [22, 45, 73]. For **iteration**, direct manipulation goes beyond text via sketch/region/layer editing with previews and before–after differences [14, 17], accessibility-oriented selection for blind/low-vision creators [32], and compositional block/voxel workflows [66, 70]. For **scale**, interfaces provide parameter sweeps [48], clustering and spatial layouts [7], diagrammatic re-organization [35], and dimension-aware/multi-level exploration [74–76]. Other directions blend retrieval and editing [72], support reference recombination [15], and enable spreadsheet-like prompting [2] or many-output sensemaking [26]. Toolchains increasingly support replayable strategy and mixed-initiative control [18, 81].

2.3 Designerly Lenses: Fidelity, Dimensionality, and Structured Variation

Stage fidelity begins with low-fidelity outputs to keep exploration inexpensive and critique candid, then increases specificity as intent sharpens [11, 67, 80]. **Dimensionality** emphasizes making orthogonal factors nameable and navigable so users can reason about cause–effect changes [49]; interfaces can expose axes explicitly [74, 76] and stabilize

shared vocabulary via terminology scaffolds [19, 41, 69]. **Structured variation** encourages one-factor-at-a-time near-miss sets that sharpen perceived differences [49]; parallel prototyping improves exploration and downstream decisions [20, 21, 77]. A constraint-based perspective clarifies how tools shape operative spaces and traversal strategies [5].

2.4 Comparison and Collaborative Sensemaking

Aligned, like-with-like comparison foregrounds structural matches and improves judgment [25]. Side-by-side evaluation surfaces trade-offs and shifts preferences in predictable ways [31, 60]; normalized grids reduce cognitive load [59, 60]. External representations support offloading and resumption over time [1, 57, 68, 85]. Artifact-anchored discussion on shared displays helps teams notice and name differences and build common ground [28, 84]; gallery-style canvases enable forking, remixing, and convergence on commensurate representations [13]. Asynchronously, teams need coordination/awareness [27, 84], structured comparability [13, 59], between-session reflection [33, 82], and externalized rationale [79, 85]. Preserving decision trails and standardizing representations lowers resumption costs [36, 52, 64], and surfacing disagreement supports multi-criteria decisions [46]. Classic rationale schemes (e.g., QOC) couple options and criteria to make reasoning inspectable [47].

User Need Synthesis:

Teams need *commensurate representations* that make alternatives comparable and trade-offs explicit; *portable rationale*—labels and traces that carry intent; *dimensional scaffolds* to help novices notice/name what to vary; and *structured variation with aligned comparison* to preserve diversity while converging. These needs motivate our canvas (§3); Appendix D details the needs–interventions mapping.

3 Tool Design: Dimensions, Differences, and Fidelity

VariationWeaver embodies three designerly principles: (a) *dimensional scaffolding* that makes salient factors explicit and reusable; (b) *intentional variation with aligned, side-by-side comparison* so differences are legible and choices portable; and (c) *staged fidelity* so output detail matches the specificity of the designer's intent.

3.1 Design Goals

DG1 — Make dimensions and terminology explicit and durable. Surface factors that matter, provide brief terms and definitions, and keep labels stable so rationale can travel across iterations and handoffs.

DG2 — Support intentional variation and aligned comparison. Enable controlled one-factor changes against a stable background, and present alternatives in commensurate side-by-side views to make trade-offs explicit.

DG3 — Stage fidelity to match design intent. Begin sketch-like to invite broad exploration and candid critique; increase fidelity as intent sharpens so evaluative detail is preserved when needed.

3.2 User Experience

Layout and workflow. VariationWeaver brings together a *Design Sidebar* for prompting and terminology and an infinite, zoomable *Image Canvas* for organizing output (Figs. 2–3). A short brief seeds a palette of named dimensions and example tags. Designers iterate three moves:

Vary. Non-target dimensions are frozen while one factor is changed to generate near-miss sets. Each image card records its provenance—prompt, active tags, freeze mask, sweep parameters, and parent—so runs can be inspected and reproduced. This makes differences attributable and supports DG2.

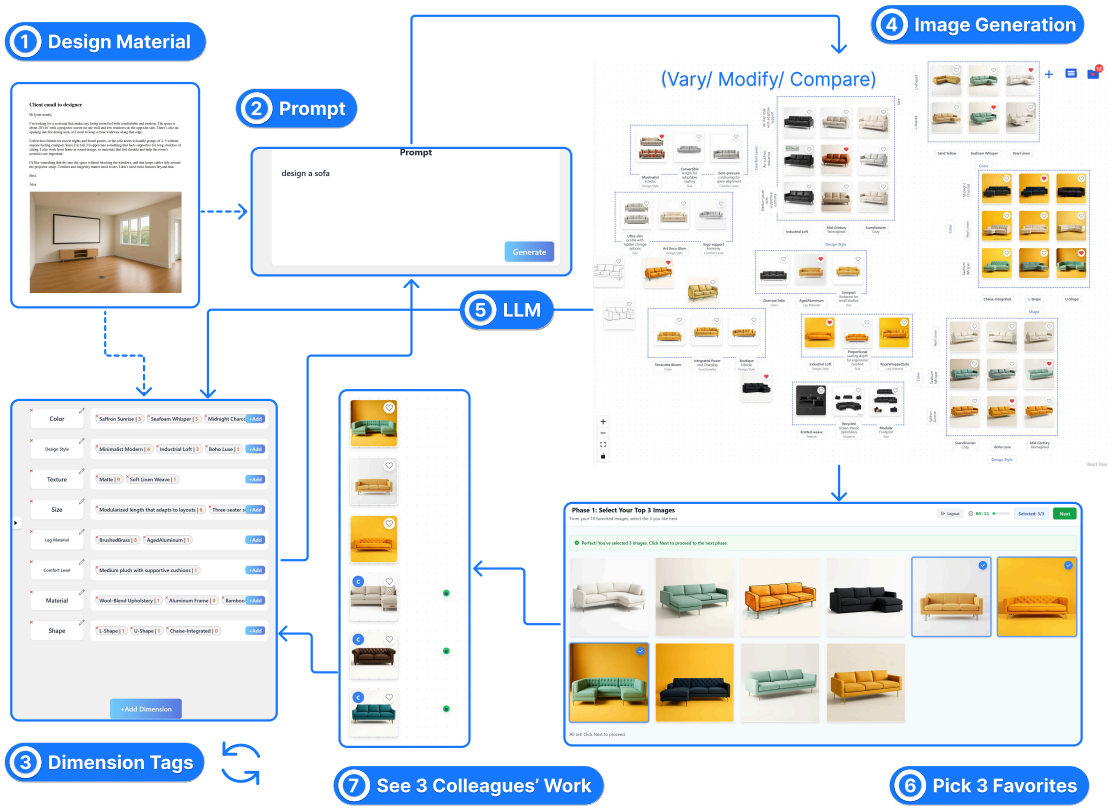

Fig. 2. Overview of *VariationWeaver*. The figure illustrates the system workflow with seven main components: (1) Design Material provides the base input for exploration; (2) Prompt allows users to specify or refine text instructions; (3) Image Generation (Vary/Modify/Compare) displays generated outputs and supports variation, modification, and comparison; (4) Dimension Tags organize structured aspects of the design space; (5) LLM analyzes photos to suggest additional tags and dimensions that enrich design exploration; (6) Pick 3 Favorite lets users curate their own preferred designs for final selection; and (7) See 3 Colleagues' Work enables viewing peer selections for collaboration.

Compare. Candidates can be arranged into an aligned matrix of dimension by value. Designers can pin a baseline, collapse near-duplicates, and attach a short criterion and rationale to the chosen value. Side-by-side inspection makes correspondences visible and articulates trade-offs, advancing DG2 and producing commensurate artifacts for handoff.

Refine terminology. Dimensions and tags are editable in place. Accepted suggestions from an image-grounded analysis fold back into the palette, keeping vocabulary synchronized across prompts, labels, and comparisons. Renames propagate automatically, supporting DG1.

Fidelity is staged across the flow: early exploration uses faster, sketch-like previews that foreground structure; later selections render at higher fidelity to support judgment (DG3). The result is a canvas where designers can move fluidly between overview and detail, and where choices and their rationale are lightweight to capture.

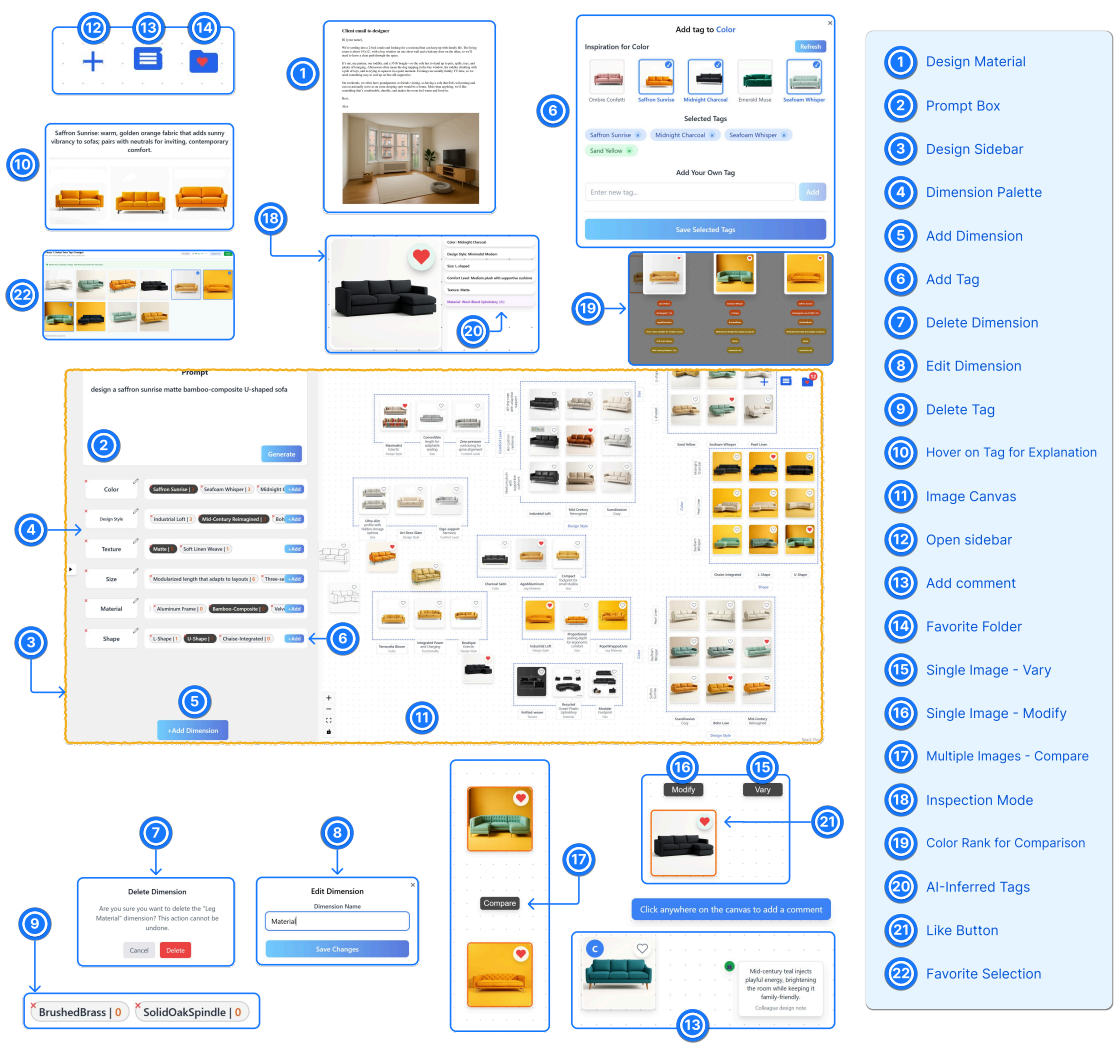

Fig. 3. User Interface of VariationWeaver. The UI facilitates structured dimensional tagging and interactive exploration of AI-generated designs. Key features include a prompt box for input, a dimension palette for organizing and modifying design aspects, and an image panel displaying generated outputs. Users can add or delete dimensions, tag designs, view detailed image information, and, with a commenting system to enable asynchronous collaboration, curate favorite designs for final selection.

3.3 Implementation Details

System stack. The front end is built in React and manages the prompt box, dimension palette, vary/compare interactions, favorites, comments, and the infinite canvas (React Flow). A Node/Express back end coordinates model calls and persistence. Firestore stores structured metadata (prompts, tags, freeze masks, parameters, selections, notes, palette state) and Firebase Storage stores images. Event logs capture page entry/exit and key interactions.

Models and policies. A lightweight language–vision model supports prompt synthesis, terminology cleanup, and image-grounded tag inference. Image generation uses a faster backend for previews (512×512) and a higher-fidelity

backend for finals (1024×1024), aligning with the staged-fidelity policy. During comparison, an axis-ordering heuristic surfaces dimensions with greater dispersion within the current set to highlight informative contrasts.

Reproducibility and release. Appendix A details prompts, parameters, and back ends; Appendix B provides design materials and the standardized simulated-colleague packet. Code and documentation will be released publicly upon publication (links withheld for review).

4 USER STUDY

We conducted a qualitative lab study to examine how *VariationWeaver* supports novice product ideation and convergence in an *asynchronous-by-proxy* setting. Rather than benchmarking against another tool, we sought interaction patterns, breakdowns, and early learning signals to guide subsequent system iteration.

4.1 Design Task and Materials

We chose a sofa design task because most people have direct experience with sofas, providing accessible common ground, yet the domain still involves meaningful trade-offs (e.g., silhouette, arm profile, leg geometry, upholstery, materials). To *diversify context without changing task difficulty*, participants were randomly assigned to **one of two briefs** (Appendix B.1). Both briefs asked participants to design a sofa for a client named *Alex* and provided equivalent constraints, but the persona and setting were intentionally varied (e.g., demographics, household composition, job context, city, and room geometry). This design broadens coverage and reduces overfitting to a single scenario while holding the core task constant. Each packet contained (i) a short client email, (ii) a 30-minute designer persona intake, and (iii) a room image to establish spatial constraints.

We used a two-stage structure to separate within-person exploration from cross-artifact comparison and to approximate asynchronous collaboration. In *Stage 1 (individual exploration)*, participants read the brief, used the canvas to create variations, and curated a top-three set. For each top pick, they wrote a one-sentence rationale that referenced salient *dimension(s)* or criteria (e.g., comfort, fit, material). In *Stage 2 (asynchronous handoff by proxy)*, participants received a standardized *simulated colleague* packet—three system-generated candidates with one-sentence notes representing another designer's interim thinking (Appendix B.2). Participants compared these against their own picks to produce a final shortlist (up to three), again providing brief, dimension-referenced rationales. This sequencing lets us observe how novices first orient to the space on their own and then negotiate trade-offs when facing another person's labeled alternatives.

4.2 Participants

We recruited **15** novice designers from a university subject pool (credit compensation). Sessions followed an IRB-approved protocol with informed consent; audio/video were recorded and de-identified for analysis. Prior exposure is summarized descriptively in Appendix E (Fig. 8): most participants reported little to no text-to-image experience, while LLM use ranged from infrequent to daily.

4.3 Procedure

Each in-lab session (~60 minutes) comprised: (1) a pre-study background survey; (2) a short tutorial and think-aloud instructions; (3) *Stage 1* creation (generate alternatives; mark favorites; write one-sentence rationales); (4) a brief in-study probe; (5) *Stage 2* handoff (review the simulated colleague's three designs and notes; compare to one's favorites;

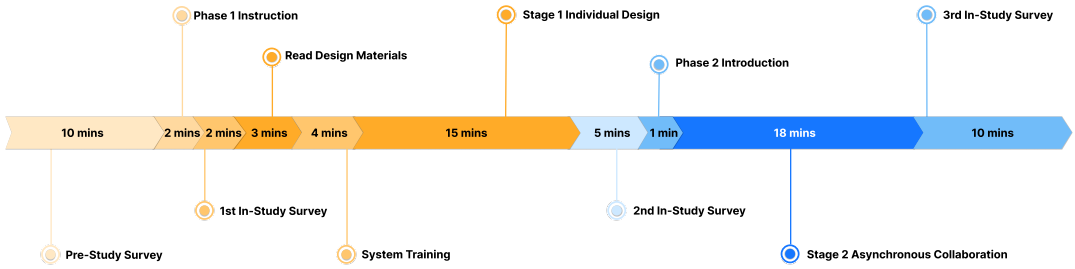

Fig. 4. Study session timeline: pre-survey, tutorial, Stage 1 (individual exploration; top-3 + rationales), probe, Stage 2 (simulated handoff; compare; shortlist + rationales), post-survey and interview.

finalize a shortlist with rationales); and (6) a post-study questionnaire followed by a short semi-structured interview. Think-aloud audio and screen capture ran throughout. A compact timeline appears in Figure 4.

4.4 Data Collection and Analysis

We analyzed think-aloud and interview *audio/transcripts* collected pre-, mid-, and post-session (Appendix C). Three researchers conducted a reflexive thematic analysis following Braun and Clarke [8, 9]: familiarization and memoing; independent open coding on a stratified subset; collaborative refinement of code meanings; primary-analyst coding of the full corpus with iterative updates; and team theme construction with attention to negative cases and meaning saturation [29, 50]. Critical-friend peer debriefs [78] followed HCI-methods guidance [6, 40]. Quantitative items are reported descriptively to contextualize the qualitative themes (Appendix E).

5 FINDINGS

We report qualitative themes from a reflexive thematic analysis of think-aloud sessions, interviews, artifacts, and logs. Findings are organized by our research questions: **RQ1** on how language support shaped dimensional acclimation; **RQ2** on how variation, comparison, and fidelity supported exploration and choice; and **RQ3** on participants' experience (confidence and ownership). A ranking probe provides descriptive context (Fig. 5); these counts are not inferential.

5.1 RQ1: Dimensional acclimation—how language support shaped sensemaking

Theme 1: Dimensional acclimation. Novices learned to navigate the design space by linking unfamiliar terms to visuals, using examples to anchor intent, and iterating with tags. Embedded terminology thus served both as vocabulary and as an acclimation medium.

T1.1 Learning terminology through visual scaffolds. Participants picked up design terms when words were paired with previews or descriptors. P1 explained, "*When I hover a tag and see a little picture, I know what it means—that's how I learn the language.*" Such scaffolds helped connect domain language to recognizable features and broaden scope (P2: "*I didn't think about durability at first, but seeing it later made me add it*"), while several admitted vocabulary gaps without these supports (P1: "*I don't know enough of the vocab…I'd describe it in simpler terms*").

T1.2 Anchoring intent with preset exemplars. Examples and labeled images stabilized participants' intent when words were hard to find. P5 found that "*using the inspiration photos gave a more varied response than typing, because the prompts*

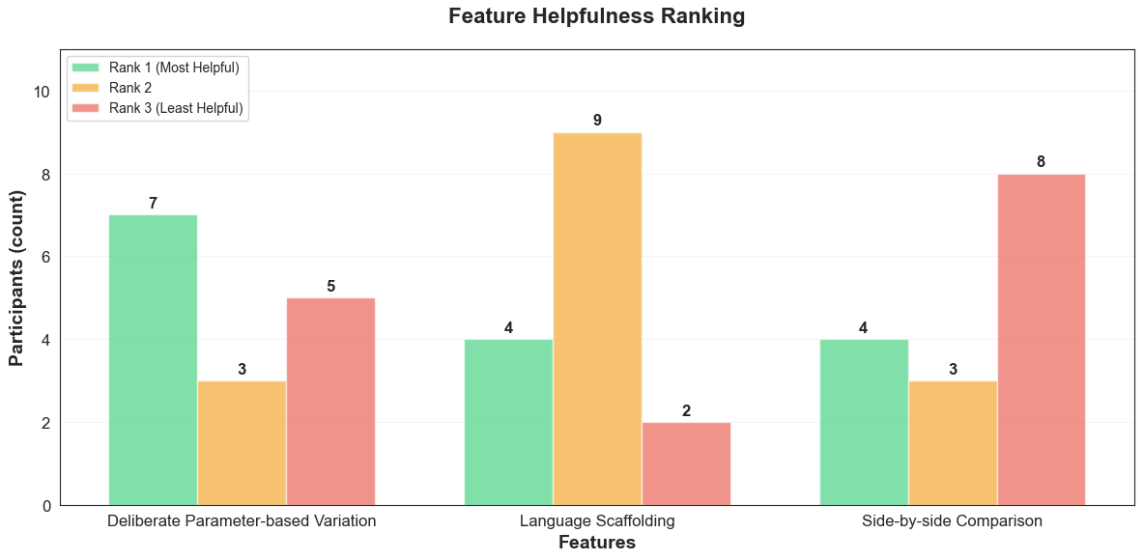

Fig. 5. Self-reported helpfulness ranking (N=15). Variation received the most Rank-1 votes; language scaffolding most Rank-2; comparison most Rank-3.

were so detailed," while P8 noted that exemplars helped fill knowledge gaps: "*Even if I didn't know what material couches were made of, [the inspiration section] gave me an idea.*" Seeing tags mapped back onto images also taught usable terms (P1: "*The labeled images taught me new words; once it was on screen, the idea felt more concrete*").

T1.3 Acclimating through tag-based refinements. By iterating on tags, novices learned how to adjust designs in ways that felt controllable. Vague edits stalled progress, while specific or dimensional changes unlocked variety (P5: "*Subtle changes…the model wouldn't pick up…I started getting more specific or pivoting to new dimensions*"). Others wanted stronger levers for big moves (P3: "*If I said large and tall, make the couch dramatically larger, and then I could scale from there*"). Over time, tags became a manageable vocabulary for tweaking outputs (P12: "*Being able to change the tags instead of everything makes it easier to tweak without changing the whole design*").

5.2 RQ2: Comparison, and Fidelity—how novices explored and decided

Theme 2: Side-by-side comparison. Comparison was designed to help novices contrast two or more images by surfacing shared and differing tags. Participants diverged sharply in how they valued this feature. A small group (4/15) found comparison the most helpful, while most others (8/15) ranked it least helpful (Fig. 5).

T2.1 Comparison valued when differences were clear. Some (7/15) reported that comparison clarified differences and guided decisions. P2 explained, "*I like the compare tool, it lets you see all the differences between the sofas and you can fill in the empty tags*," while P10 said it was "*easy to inspect what was missing when two images were compared.*" P10 described functional contrasts (e.g., cupholder, recliner). A few ranked comparison as their top feature: P4 reflected, "*the comparative interface was number one for me, because I didn't feel like I was able to bring about the sofa I wanted until I looked at each sofa's dimensions side by side.*" Interaction records also show active use (e.g., P15 merged images and generated matrices to explore subtle differences).

T2.2 Comparison weakened when differences were trivial or unclear. For others (8/15), comparison felt redundant, especially when images looked similar or surfaced differences were ones they had already manipulated. As P14 put it, "*the least helpful would be a comparative interface… a lot of the designs I made were really similar… usually only one difference, and it was one I had already manipulated.*" Beyond redundancy, participants struggled when differences were not visually salient. P3 explained they had to "*guess the dimensions… because sometimes the AI didn't respond*," and wished for preset tags to guide comparisons. P4 similarly noted that style and size labels like *modular* or *chute* were unclear, so they relied on obvious properties such as color and height. P14 echoed this, emphasizing, "*I'm a very visual person, so just looking at how they were visually different was impactful.*"

Theme 3: Staged fidelity. The staged fidelity of outputs, starting with sketches and moving toward polished renders, shaped how participants generated and evaluated ideas. For some participants (6/15), sketch-like renderings were generative, sparking creativity and leaving space for interpretation. For most others (9/15), sketches made evaluation harder, leading to confusion or frustration until polished images arrived.

T3.1 Sketches as support for generativity and flexibility. Six participants appreciated sketches for sparking ideas and leaving space for flexibility. P6 reflected that "*maybe the sketches… make me to create the product from scratch myself, and I have more broad idea to how to make it.*" P8 similarly noted that starting from a rough sketch "*was good… because I was like, oh wow, that's really ugly, I need to get creative to make it look better… it forced me to… make this better than what it is, and make it fit.*" Logs show direct iteration on sketches (P2, P3, P12). As P14 explained, "*a rough idea is better… it gives me a better visualization… before I think about more specific features*," and P2 added that sketch form is helpful because "*you'll like to know what's modified before you get… the final render… in case you wanna change stuff around.*"

T3.2 Sketches as obstacles to clarity. At the same time, a majority (9/15) felt sketches reduced clarity and complicated evaluation. P4 explained, "*when it said sofa and it gave me a drawing, that was very annoying… I thought it would give me an actual sofa, not a 2D sketch… I didn't feel like I could tell what I was looking at*," while P11 similarly said, "*when you just have a sketch… I can't really imagine that in the space that I have with a lot of… clarity.*" Others emphasized that polished outputs made evaluation easier: P9 stated, "*I liked seeing the polished, finished products… it just gives me something to work with, like ideas. Because I'm not really creative without some inspiration*," and P5 highlighted that a finished product "*looked really realistic… I could see clearly what I would want to change.*" P15 recalled being "*thrown… it was a little bit disorienting*," but noted that "*once they were polished after the fact… it made it feel like it was more real.*" Logs echoed these struggles (e.g., P1 abandoned base sketches; P3 noted repeated sketch iterations failed to incorporate intended tweaks).

5.3 RQ3: User Experience—confidence and ownership while designing with the canvas

Theme 4: Confidence shifts. When reflecting on their experience, participants varied in whether and how their confidence in designing changed. About half (7/15) reported an increase, citing support from the GenAI workflow or insights from colleagues' designs; others described stability or modest boosts tied to tool fluency rather than deeper changes.

T4.1 Confidence gained from the GenAI workflow. Four participants felt more confident because the workflow clarified next steps and reduced uncertainty. For example, P4: "*it was easier using AI… made it easier for me*," while P3 reflected "*more confident… now I know what the task requires.*" Several noted that adding one tag at a time and seeing immediate results made progress feel achievable.

T4.2 Confidence gained from colleagues' design. Three participants credited confidence increases to seeing and working with colleagues' outputs. P6: "*it's shift… to a better level… I got more ideas from the colleague, and I could merge them together.*" P10 emphasized reduced pressure: "*less pressure on me… I was able to use some of theirs and create a sofa.*" P12 added, "*a little bit… because I noticed that some of the other people's designs were similar to mine.*"

T4.3 Confidence steady or tied to tool fluency. Eight reported no change or small boosts tied to practice. P7: "*it stayed the same, because… it's the same thing*," and P4: "*has not shifted… I remain moderately confident.*" Others noted comfort from practice, e.g., P11: "*honestly, fairly confident… I got the hang of the AI tool… shifted from moderate to fairly*," and P12: "*a little more confident now that I got to practice.*"

Theme 5: Ownership. Alongside confidence, we examined how novices described ownership of the outputs. Roughly half (7/15) said the results felt like their designs once outputs matched their ideas or could be shaped through edits; the rest (8/15) reported limited or absent ownership.

T5.1 Ownership emerged when outputs matched or could be shaped to participants' ideas. Several said the results felt like "their design" once the system produced something close to their intent. P2: "*this is exactly what I wanted… I felt like my concept rather than something produced by the AI… I tried to get it to follow what I was thinking instead of doing its own thing.*" P3: "*at the end, when the AI finally responded to what I was looking for… seeing the change I was typing being made felt like my design.*" Others described ownership building through refinements. P11: "*when you start making all of those modifications—shape, width, functionality, and color—it starts feeling more like your idea instead of just AI.*" P14: "*nearer to the end… it was a lot of filtering of what I thought was important, and that made it more personal.*" Curating among alternatives also helped (P15: "*a combination of both… I had more say when I had multiple concepts in front of me.*")

T5.2 Ownership was limited when outputs stayed distant from intentions. The remaining participants either never felt the sofas were their own or only felt some ownership very late. P5: "*overall… didn't feel like my concept, my influence was smaller than the AI's influence*," while P4: "*never really felt like my idea… only at the very, very end when I finally specified a three-seater sofa and got the image I had in mind.*" For others, images seemed detached from intentions. P9: "*none of it… the images looked super AI-generated and not what I was trying to create.*" In these cases, ownership emerged only after trial and error—or not at all.

6 DISCUSSION

We interpret the mixed patterns in §5, explaining why visible language, comparison, and staged fidelity produced both benefits and friction. We connect to prior work and outline focused future work.

6.1 Why visual language scaffolds accelerated acclimation

Participants learned fastest when terms were *visible in place* and grounded in visuals (§5.1–5.1). This aligns with Information Foraging Theory: strong "scent" (clear labels with immediate payoffs) reduces navigation cost and improves learning [54, 55, 61]. Micro-previews plus tags also leverage external representations to offload cognition and help map terminology to perceptual features [1, 68]. Our positive cases (P1, P2, P5, P8, P12) follow directly: previews suggested what to vary, exemplars anchored intent, and tag edits created a safe, incremental control surface. Where labels were ambiguous or missing, people reverted to vaguer prompting—consistent with weaker scent and higher search cost.

Design implication. Treat the palette as a learnable, durable lexicon: pair each term with a tiny exemplar, keep terms stable across turns, and support image→tag remapping so users can "pull" language from what they see. This matches both the evidence and theory.

6.2 When comparison helps—and when it hinders

Comparison helped when alternatives were *commensurate* and differences *diagnostic* (§5.2). Structural alignment predicts that aligned formats make correspondences legible [25], while joint, side-by-side evaluation elicits trade-offs and reduces memory load [31, 59, 60]. Participants' positive accounts (P2, P4, P10, P15) mirror this: clear axes, visible diffs, and actionable contrasts.

By contrast, comparison faltered when sets were too homogeneous, when labels were opaque, or when the model ignored targeted edits (§5.2). Variation Theory holds that without *systematic contrast*—varying one factor while others remain invariant—learners cannot readily discern critical features [49]. Non-responsiveness collapses near-miss structure; the matrix then surfaces little that users did not already know, leading to redundancy (P14) or guesswork (P3). Ambiguous terms further break alignment, shifting people to purely visual scanning (P4, P14).

Design implication. Engineer near-miss sets (freeze non-target axes), delay comparison until dispersion is sufficient, and overlay explicit change badges (e.g., *Arm: track → pillow*). Order axes by dispersion to foreground informative contrasts [60]. These steps generalize directly from the successful patterns we observed.

6.3 Staging fidelity for exploration *and* evaluation

Participants split on sketch-like outputs (§5.2). Prototyping research explains the tension: low fidelity reduces commitment and invites broad exploration; high fidelity supports specification and judgment [11, 67, 80]. Our data show both sides. For some (P6, P8, P14), sketches promoted lateral thinking and incremental refinement; for others (P4, P11, P15), sketches obscured evaluative cues until polished images arrived. Additionally, when successive sketches failed to reflect targeted edits (P1, P3), users questioned whether the system respected their intentions.

Design implication. Couple fidelity not only to prompt specificity but also to user control and task intent. Make fidelity *visible, adjustable, and reversible*; keep side-by-side provenance so users can verify that only intended dimensions changed. This reconciles sketch benefits with the need for clarity at decision time.

6.4 Confidence and ownership: scaffolds, social cues, and agency

Confidence increased when the workflow scaffolded next steps (P3, P4) or when colleague packets provided labeled exemplars (P6, P10, P12; §5.3). Seeing multiple aligned alternatives can foster integration and rapport [20, 21], consistent with reports that social exposure reduced pressure and validated direction. Ownership rose when participants saw their input reflected or could shape results through small, visible moves (§5.3), aligning with benefits of incremental, inspectable change for maintaining agency. Conversely, ownership eroded when the model ignored targeted edits (P5, P9), undercutting both comparison value (no reliable diffs) and perceived authorship—again consistent with the role of systematic contrast in learning and control [49].

6.5 Limitations

Our lab study focuses on novices (N=15) and one domain (sofas), which may limit generality. The *asynchronous-by-proxy* design approximates, but does not replicate, multi-party handoffs. We did not include a comparative baseline; the

ranking probe is descriptive. Model non-responsiveness sometimes broke intended near-miss structure, confounding comparison. Finally, we did not directly measure fixation or coverage; mechanistic claims are interpretive.

6.6 Future Work

Image-centric variation and comparison. Findings suggest novices perceive diffs best via visual contrast and one-factor changes. We will prototype row/strip generators, explicit "what changed" badges, and dispersion-aware axis ordering, then ablate components to estimate contributions to search breadth and time-to-decision.

Multi-person, asynchronous collaboration. Extend beyond a proxy to small teams: per-dimension locks for partial agreement, branch-and-merge on variation sets, and criterion-linked comments that travel with artifacts. Longitudinal deployments can test how shared terminology stabilizes and what provenance granularity supports accountable convergence [28, 85].

Fidelity controls and evaluation. Experiment with user-visible fidelity policies (sketch-first vs. render-first vs. user choice) and quantify effects on breadth, difference noticing, and convergence quality [67, 80].

7 CONCLUSION

VariationWeaver shows how making dimensions visible, varying one factor at a time with aligned comparison, and staging fidelity can help novices explore and converge with text-to-image models. In a qualitative study with 15 participants, terminology scaffolds accelerated dimensional acclimation; variation and side-by-side comparison supported deliberate trade-offs when differences were diagnostic; and sketch-first outputs encouraged lateral exploration but sometimes impeded evaluation—clarifying when fidelity should rise. Together, these results argue for commensurate, artifact-anchored views and durable vocabulary that move work from prompting toward designerly reasoning. We outline next steps toward image-centric diff workflows, multi-person asynchronous collaboration, and controllable fidelity policies to balance breadth with clarity. More broadly, embedding variation, comparison, and staged fidelity into creative AI tools can turn fast images into teachable, decision-ready design representations.

References

[1] Shaaron Ainsworth. 2006. DeFT: A conceptual framework for considering learning with multiple representations. *Learning and instruction* 16, 3 (2006), 183–198.

[2] Shm Garanganao Almeda, J.D. Zamfirescu-Pereira, Kyu Won Kim, Pradeep Mani Rathnam, and Bjoern Hartmann. 2024. Prompting for Discovery: Flexible Sense-Making for AI Art-Making with Dreamsheets. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 160, 17 pages. doi:10.1145/3613904.3642858

[3] Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, Marjorie Limont, Matt McGill, Yanko Oliveira, Jack Parker-Holder, Frank Perbet, Guy Scully, Jeremy Shar, Stephen Spencer, Omer Tov, Ruben Villegas, Emma Wang, Jessica Yung, Cip Baetu, Jordi Berbel, David Bridson, Jake Bruce, Gavin Buttimore, Sarah Chakera, Bilva Chandra, Paul Collins, Alex Cullum, Bogdan Damoc, Vibha Dasagi, Maxime Gazeau, Charles Gbadamosi, Woohyun Han, Ed Hirst, Ashyana Kachra, Lucie Kerley, Kristian Kjems, Eva Knoepfel, Vika Koriakin, Jessica Lo, Cong Lu, Zeb Mehring, Alex Moufarek, Henna Nandwani, Valeria Oliveira, Fabio Pardo, Jane Park, Andrew Pierson, Ben Poole, Helen Ran, Tim Salimans, Manuel Sanchez, Igor Saprykin, Amy Shen, Sailesh Sidhwani, Duncan Smith, Joe Stanton, Hamish Tomlinson, Dimple Vijaykumar, Luyu Wang, Piers Wingfield, Nat Wong, Keyang Xu, Christopher Yew, Nick Young, Vadim Zubov, Douglas Eck, Dumitru Erhan, Koray Kavukcuoglu, Demis Hassabis, Zoubin Gharamani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. 2025. Genie 3: A New Frontier for World Models. (2025).

[4] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. 2024. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945* (2024).

[5] Michael Mose Biskjaer, Peter Dalsgaard, and Kim Halskov. 2014. A constraint-based understanding of design spaces. In *Proceedings of the 2014 conference on Designing interactive systems*. 453–462.

[6] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. *Qualitative HCI research: Going behind the scenes*. Morgan & Claypool Publishers.

[7] Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-Image Generation through Interactive Prompt Exploration with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 96, 14 pages. doi:10.1145/3586183.3606725

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

[9] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

[10] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 159, 25 pages.

[11] Bill Buxton. 2007. *Sketching User Experiences: Getting the Design Right and the Right Design*. Morgan Kaufmann.

[12] Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. 2023. Muse: Text-To-Image Generation via Masked Generative Transformers. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 4055–4075. https://proceedings.mlr.press/v202/chang23b.html

[13] Joseph Chee Chang, Nathan Hahn, and Aniket Kittur. 2020. Mesh: Scaffolding Comparison Tables for Online Decision Making. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 391–405. doi:10.1145/3379337.3415865

[14] Lydia B Chilton, Ecenaz Jen Ozmen, Sam H Ross, and Vivian Liu. 2021. VisiFit: Structuring Iterative Improvement for Novice Designers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 574, 14 pages. doi:10.1145/3411764.3445089

[15] DaEun Choi, Sumin Hong, Jeongeon Park, John Joon Young Chung, and Juho Kim. 2024. CreativeConnect: Supporting Reference Recombination for Graphic Design Ideation with Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1055, 25 pages. doi:10.1145/3613904.3642794

[16] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sashank Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2024. PaLM: scaling language modeling with pathways. *J. Mach. Learn. Res.*

24, 1, Article 240 (mar 2024), 113 pages.

[17] Hai Dang, Frederik Brudy, George Fitzmaurice, and Fraser Anderson. 2023. WorldSmith: Iterative and Expressive Prompting for World Building with a Generative AI. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 63, 17 pages. doi:10.1145/3586183.3606772

[18] Nicholas Davis, Xinyi Lin, Sarah Yalowitz, and Emily Walker. 2021. Supporting creative exploration in generative design tools. In *Proceedings of the 2021 ACM Conference on Creativity and Cognition*. ACM, 1–12.

[19] Giulia Di Fede, Davide Rocchesso, Steven P. Dow, and Salvatore Andolina. 2022. The Idea Machine: LLM-based Expansion, Rewriting, Combination, and Suggestion of Ideas. In *Proceedings of the 14th Conference on Creativity and Cognition* (Venice, Italy) *(C&C '22)*. Association for Computing Machinery, New York, NY, USA, 623–627. doi:10.1145/3527927.3535197

[20] Steven Dow, Julie Fortuna, Dan Schwartz, Beth Altringer, Daniel Schwartz, and Scott Klemmer. 2011. Prototyping dynamics: sharing multiple designs improves exploration, group rapport, and results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2807–2816. doi:10.1145/1978942.1979359

[21] Steven P. Dow, Alana Glassco, Jonathan Kass, Melissa Schwarz, Daniel L. Schwartz, and Scott R. Klemmer. 2011. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Trans. Comput.-Hum. Interact.* 17, 4, Article 18 (dec 2011), 24 pages. doi:10.1145/1879831.1879836

[22] Yingchaojie Feng, Xingbo Wang, Kam Kwai Wong, Sijia Wang, Yuhong Lu, Minfeng Zhu, Baicheng Wang, and Wei Chen. 2024. PromptMagician: Interactive Prompt Engineering for Text-to-Image Creation. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (Jan. 2024), 295–305. doi:10.1109/TVCG.2023.3327168

[23] Jonas Frich, Midas Nouwens, Kim Halskov, and Peter Dalsgaard. 2021. How digital tools impact convergent and divergent thinking in design ideation. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–11.

[24] Simret Araya Gebreegziabher, Yukun Yang, Elena L. Glassman, and Toby Jia-Jun Li. 2025. Supporting Co-Adaptive Machine Teaching through Human Concept Learning and Cognitive Theories. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 533, 18 pages. doi:10.1145/3706598.3713708

[25] Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science* 7, 2 (1983), 155–170.

[26] Katy Ilonka Gero, Chelse Swoopes, Ziwei Gu, Jonathan K. Kummerfeld, and Elena L. Glassman. 2024. Supporting Sensemaking of Large Language Model Outputs at Scale. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 838, 21 pages. doi:10.1145/3613904.3642139

[27] Carl Gutwin and Saul Greenberg. 2002. A descriptive framework of workspace awareness for real-time groupware. *Computer Supported Cooperative Work (CSCW)* 11, 3 (2002), 411–446.

[28] Jeffrey Heer, Fernanda B. Viégas, and Martin Wattenberg. 2007. Voyagers and voyeurs: supporting asynchronous collaborative information visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 1029–1038. doi:10.1145/1240624.1240781

[29] Monique M Hennink, Bonnie N Kaiser, and Vincent C Marconi. 2017. Code saturation versus meaning saturation: how many interviews are enough? *Qualitative health research* 27, 4 (2017), 591–608.

[30] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022).

[31] Christopher K Hsee and France Leclerc. 1998. Will products look more attractive when presented separately or together? *Journal of Consumer Research* 25, 2 (1998), 175–186.

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

[33] Shamsi T. Iqbal and Eric Horvitz. 2007. Disruption and recovery of computing tasks: field study, analysis, and directions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 677–686. doi:10.1145/1240624.1240730

[34] David G Jansson and Steven M Smith. 1991. Design fixation. *Design Studies* 12, 1 (1991), 3–11.

[35] Peiling Jiang, Jude Rayan, Steven P. Dow, and Haijun Xia. 2023. Graphologue: Exploring Large Language Model Responses with Interactive Diagrams. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 3, 20 pages. doi:10.1145/3586183.3606737

[36] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E Kraut. 2011. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. 43–52.

[37] Janin Koch, Nicolas Taffin, Andrés Lucero, and Wendy E. Mackay. 2020. SemanticCollage: Enriching Digital Mood Board Design with Semantic Labels. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS '20)*. Association for Computing Machinery, New York, NY, USA, 407–418. doi:10.1145/3357236.3395494

[38] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. 2023. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125* (2023).

[39] Harsh Kumar, Jonathan Vincentius, Ewan Jordan, and Ashton Anderson. 2025. Human Creativity in the Age of LLMs: Randomized Experiments on Divergent and Convergent Thinking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 23, 18 pages. doi:10.1145/3706598.3714198

[40] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research methods in human-computer interaction*. Morgan Kaufmann.

[41] Q. Vera Liao, Hariharan Subramonyam, Jennifer Wang, and Jennifer Wortman Vaughan. 2023. Designerly Understanding: Information Needs for Model Transparency to Support Design Ideation for AI-Powered User Experience. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 9, 21 pages. doi:10.1145/3544548.3580652

[42] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. 2023. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 300–309.

[43] Pei-Ying Lin, Kristina Andersen, Ralf Schmidt, Sanne Schoenmakers, Hèrm Hofmeyer, Pieter Pauwels, and Wijnand Ijsselsteijn. 2024. Text-to-Image AI as a Catalyst for Semantic Convergence in Creative Collaborations. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 2753–2767. doi:10.1145/3643834.3661543

[44] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. 2024. One-2-3-45++: Fast Single Image to 3D Objects with Consistent Multi-View Generation and 3D Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10072–10083.

[45] Vivian Liu and Lydia B Chilton. 2022. Design Guidelines for Prompt Engg Large Language Model Prompts through Visual Programmingineering Text-to-Image Generative Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 384, 23 pages. doi:10.1145/3491102.3501825

[46] Weichen Liu, Sijia Xiao, Jacob T. Browne, Ming Yang, and Steven P. Dow. 2018. ConsensUs: Supporting Multi-Criteria Group Decisions by Visualizing Points of Disagreement. *Trans. Soc. Comput.* 1, 1, Article 4 (jan 2018), 26 pages. doi:10.1145/3159649

[47] Allan MacLean, Richard M Young, Victoria ME Bellotti, and Thomas P Moran. 2020. Questions, options, and criteria: Elements of design space analysis. In *Design rationale*. CRC Press, 53–105.

[48] J. Marks, B. Andalman, P. A. Beardsley, W. Freeman, S. Gibson, J. Hodgins, T. Kang, B. Mirtich, H. Pfister, W. Ruml, K. Ryall, J. Seims, and S. Shieber. 1997. Design galleries: a general approach to setting parameters for computer graphics and animation. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '97)*. ACM Press/Addison-Wesley Publishing Co., USA, 389–400. doi:10.1145/258734.258887

[49] Ference Marton and Ming Fai Pang. 2006. On some necessary conditions of learning. *The Journal of the Learning sciences* 15, 2 (2006), 193–220.

[50] Lorelli S Nowell, Jill M Norris, Deborah E White, and Nancy J Moules. 2017. Thematic analysis: Striving to meet the trustworthiness criteria. *International journal of qualitative methods* 16, 1 (2017), 1609406917733847.

[51] Lora Oehlberg, Manfred Lau, and Björn Hartmann. 2012. DesignScape: Supporting creativity within UI design constraints. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1371–1380.

[52] Gary M. Olson and Judith S. Olson. 2000. Distance matters. *Hum.-Comput. Interact.* 15, 2 (Sept. 2000), 139–178. doi:10.1207/S15327051HCI1523_4

[53] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2024. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 2011, 15 pages.

[54] Srishti Palani, Zijian Ding, Stephen MacNeil, and Steven P. Dow. 2021. The "Active Search" Hypothesis: How Search Strategies Relate to Creative Learning. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (Canberra ACT, Australia) *(CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 325–329. doi:10.1145/3406522.3446046

[55] Srishti Palani, Yingyi Zhou, Sheldon Zhu, and Steven P. Dow. 2022. InterWeave: Presenting Search Suggestions in Context Scaffolds Information Search and Synthesis. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 93, 16 pages. doi:10.1145/3526113.3545696

[56] Jeongeon Park, Eun-Young Ko, Yeon Su Park, Jinyeong Yim, and Juho Kim. 2024. DynamicLabels: Supporting Informed Construction of Machine Learning Label Sets with Crowd Feedback. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 209–228. doi:10.1145/3640543.3645157

[57] Jeongeon Park, Bryan Min, Xiaojuan Ma, and Juho Kim. 2023. Choicemates: Supporting unfamiliar online decision-making with multi-agent conversational interactions. *arXiv preprint arXiv:2310.01331* (2023).

[58] Emily S Patterson, Emilie M Roth, David D Woods, Renée Chow, and José Orlando Gomes. 2004. Handoff strategies in settings with high consequences for failure: lessons for health care operations. *International journal for quality in health care* (2004), 125–132.

[59] John W Payne. 1976. Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational behavior and human performance* 16, 2 (1976), 366–387.

[60] John W Payne, James R Bettman, and Eric J Johnson. 1993. *The adaptive decision maker*. Cambridge university press.

[61] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review* 106, 4 (1999), 643.

[62] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988* (2022).

[63] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8821–8831. https://proceedings.mlr.press/v139/ramesh21a.html

[64] Daniela Retelny, Sébastien Robaszkiewicz, Alexandra To, Walter S. Lasecki, Jay Patel, Negar Rahmati, Tulsee Doshi, Melissa Valentine, and Michael S. Bernstein. 2014. Expert crowdsourcing with flash teams. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 75–85. doi:10.1145/2642918.2647409

[65] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

[66] Vishnu Sarukkai, Lu Yuan, Mia Tang, Maneesh Agrawala, and Kayvon Fatahalian. 2024. Block and Detail: Scaffolding Sketch-to-Image Generation. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 33, 13 pages. doi:10.1145/3654777.3676444

[67] Juergen Sauer and Andreas Sonderegger. 2009. The influence of prototype fidelity and aesthetics of design in usability tests: Effects on user behaviour, subjective evaluation and emotion. *Applied ergonomics* 40, 4 (2009), 670–677.

[68] Mike Scaife and Yvonne Rogers. 1996. External cognition: how do graphical representations work? *International journal of human-computer studies* 45, 2 (1996), 185–213.

[69] Orit Shaer, Angelora Cooper, Osnat Mokryn, Andrew L Kun, and Hagit Ben Shoshan. 2024. AI-Augmented Brainwriting: Investigating the use of LLMs in group ideation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1050, 17 pages. doi:10.1145/3613904.3642414

[70] Xinyu Shi, Yinghou Wang, Ryan Rossi, and Jian Zhao. 2025. Brickify: Enabling Expressive Design Intent Specification through Direct Manipulation on Design Tokens. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 424, 20 pages. doi:10.1145/3706598.3714087

[71] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792* (2022).

[72] Kihoon Son, DaEun Choi, Tae Soo Kim, Young-Ho Kim, and Juho Kim. 2024. GenQuery: Supporting Expressive Visual Search with Generative Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 180, 19 pages. doi:10.1145/3613904.3642847

[73] Hari Subramonyam, Roy Pea, Christopher Pondoc, Maneesh Agrawala, and Colleen Seifert. 2024. Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1039, 19 pages. doi:10.1145/3613904.3642754

[74] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 644, 26 pages. doi:10.1145/3613904.3642400

[75] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 1, 18 pages. doi:10.1145/3586183.3606756

[76] Sirui Tao, Ivan Liang, Cindy Peng, Zhiqing Wang, Srishti Palani, and Steven P. Dow. 2025. DesignWeaver: Dimensional Scaffolding for Text-to-Image Product Design. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 425, 26 pages. doi:10.1145/3706598.3714211

[77] Maryam Tohidi, William Buxton, Ronald Baecker, and Abigail Sellen. 2006. Getting the right design and the design right. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI '06)*. Association for Computing Machinery, New York, NY, USA, 1243–1252. doi:10.1145/1124772.1124960

[78] Sarah J Tracy. 2010. Qualitative quality: Eight "big-tent" criteria for excellent qualitative research. *Qualitative inquiry* 16, 10 (2010), 837–851.

[79] Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. 2004. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vienna, Austria) *(CHI '04)*. Association for Computing Machinery, New York, NY, USA, 575–582. doi:10.1145/985692.985765

[80] Miriam Walker, Leila Takayama, and James A Landay. 2002. High-fidelity or low-fidelity, paper or computer? Choosing attributes when testing web prototypes. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 46. Sage Publications Sage CA: Los Angeles, CA, 661–665.

[81] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 385, 22 pages. doi:10.1145/3491102.3517582

[82] Xiaotong (Tone) Xu, Jiayu Yin, Catherine Gu, Jenny Mar, Sydney Zhang, Jane L. E, and Steven P. Dow. 2024. Jamplate: Exploring LLM-Enhanced Templates for Idea Reflection. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 907–921. doi:10.1145/3640543.3645196

[83] S. Yilmaz, S.R. Daly, C.M. Seifert, and R. Gonzalez. 2015. Design heuristics in innovative products. *Journal of Mechanical Design* 137, 7 (2015), 071102.

[84] Amy X. Zhang, Lea Verou, and David Karger. 2017. Wikum: Bridging Discussion Forums and Wikis Using Recursive Summarization. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2082–2096. doi:10.1145/2998181.2998235

[85] Jiaje Zhang and Donald A Norman. 1994. Representations in distributed cognitive tasks. *Cognitive science* 18, 1 (1994), 87–122.

A PROMPTS AND MODEL PIPELINES

We standardize model names as `gpt-5-nano` (language / lightweight vision), `gpt-image-1` (image generation and image-to-image), and `flux-schnell` (fast sweeps). JSON shown is schematic.

A.1 Prompt Synthesis from Tags

- **Goal:** Convert active dimension tags into a short, natural prompt.
- **Model(s):** `gpt-5-nano`.
- **Messages: user** — "Create a simple sofa design prompt with these characteristics: {`tagDescription`}. Keep it simple and direct. Return only the prompt."
- **Inputs:** `tagDescription` (NL phrase from active tags).
- **Outputs:** Prompt string (quotes stripped), e.g., `design a black leather modern sofa`.

A.2 Image Generation Backends & Styles

- **Goal:** Render images for browsing, sweeps, and comparison.
- **Model(s):** `gpt-image-1` (1024×1024); `flux-schnell` (512×512).
- **Messages:** Text prompts composed from user text + tag clauses; common tail enforces one sofa, white background, studio lighting.
- **Inputs:** Prompt string; optional reference image(s) for image-to-image; optional seed token.
- **Outputs:** PNG/JPEG (base64 → Firebase URL). Style ramp: 0–2 tags (BW sketch), 3–4 (color sketch), ≥5 (studio product).

A.3 Controlled Variation (Sweep) Prompts

- **Goal:** Generate an $m \times n$ grid by freezing non-target axes and sweeping chosen values on one or more target dimensions.
- **Model(s):** `gpt-image-1` (prefer image-to-image); fallback `flux-schnell`/text-to-image.
- **Messages: user** — "CRITICAL INSTRUCTION: change $\texttt{Dimension}_i$ `to` $\texttt{Value}_i$ [...] in the original image; keep angle/background/lighting/style fixed. (`tagDescription`). Variation `#i/K`. Use identifier `seed_...`."
- **Inputs:** Original image URL; frozen & target axes; value lists; user prompt; tagDescription.
- **Outputs:** Grid of images + per-cell metadata (prompt, tags, freeze mask, sweep params, parentId).

A.4 Image-grounded Tag Inference

- **Goal:** Fill missing dimension–tag pairs or propose complementary ones from an image.
- **Model(s):** `gpt-5-nano` (vision input).
- **Messages:**

  - **Focused fill** — **system**: "ONLY infer {`focusDimensions`}; return EXACT JSON of that size." **user**: existing tags + image.
  - **Open-ended complement** — **user**: request k non-overlapping, visually salient pairs; return JSON.
- **Inputs:** Image (data URL); `existingTags`; optional `focusDimensions` or k.
- **Outputs:** JSON object of new dimension→tag pairs (title-cased keys, verbatim values).

A.5 Tag Recommendation (within a Dimension)

- **Goal:** Suggest diverse tag values within a chosen dimension.
- **Model(s):** `gpt-5-nano`.
- **Messages: user** — "For dimension '{`dimensionKey`}', generate 5 diverse, new tags. Return a JSON array."
- **Inputs:** `dimensionKey`; optional brief context.
- **Outputs:** `[tag_1, . . . , tag_5]` (unique, non-duplicate).

A.6 Dimension Recommendation (new axes)

- **Goal:** Propose new dimensions plus one canonical tag each.
- **Model(s):** `gpt-5-nano`.
- **Messages: system** — "Generate exactly {`targetCount`} NEW dimensions; one tag each; ONLY JSON." **user** — short description + list of dimensions to avoid.
- **Inputs:** Prompt text; `existingDimensions`; `targetCount`.
- **Outputs:** JSON object {`Dimension: Tag`}$^{\times \texttt{targetCount}}$.

A.7 Terminology Propagation & Normalization

- **Goal:** Keep names/values consistent across UI, prompts, and comparisons after edits.
- **Model(s):** None (deterministic transforms).
- **Messages:** N/A.
- **Inputs:** User edits (add/rename/remove); current palette; selected tags.
- **Outputs:** Updated palette; synchronized labels in prompts, grids, and compare view (case-insensitive key replace; title-cased keys; verbatim values).

A.8 Prompt Cleaning

- **Goal:** Remove artifacts (dangling "with", stray commas, double spaces) while preserving user phrasing.
- **Model(s):** `gpt-5-nano` with regex fallback.
- **Messages: system** — "Return only the cleaned prompt." **user** — "Clean this prompt: "{`prompt`}"."
- **Inputs:** Raw prompt string.
- **Outputs:** Cleaned prompt string (or regex-cleaned fallback).

A.9 Data & Logging

- **Goal:** Persist artifacts and lightweight analytics.

- **Model(s):** None.
- **Messages:** N/A.
- **Inputs:** Prompts, tags, sweeps, selections, notes, palette state; page entry/exit timestamps.
- **Outputs:** Firestore docs (structured metadata); Firebase Storage image URLs; session duration summaries (no sensitive telemetry).

A.10 Axis Ranking & Similarity (Compare View)

- **Goal:** Rank informative axes and order grids.
- **Model(s):** None (heuristics).
- **Messages:** N/A.
- **Inputs:** Candidate sets with per-image tags.
- **Outputs:** Axis scores (Jaccard dispersion on tag sets; optional Levenshtein); ordering for compare view.

A.11 Compare-mode Generative Operations (Optional)

- **Goal:** Synthesize a child candidate by blending two or more references or enforcing a chosen `dimension:value`.
- **Model(s):** `gpt-image-1` (primary); `flux-schnell` (previews/fallback).
- **Messages:** Prompt skeleton instructs blend of parent tag sets and enforcement of selected `dimension:value`; common tail (single sofa, white background).
- **Inputs:** Reference image URLs; `tags1`, `tags2` (or list); chosen `dimension:value`; optional user prompt.
- **Outputs:** Child image URL; merged tags (dimension-wise merge with enforced value).

B Design Materials & Simulated Colleague

B.1 Design Materials

Version A.

Client email: Hi [your name], We're settling into a 2-bed condo and looking for a sectional that can keep up with family life. The living room is about $14' \times 12'$*, with a bay window on one short wall and a balcony door on the other, so we'll need to leave a clear path through the space. It's me, my partner, our toddler, and a 35-lb beagle—so the sofa has to stand up to pets, spills, toys, and plenty of lounging. Afternoons often mean the dog napping in the bay window, the toddler climbing with a pile of toys, and us trying to squeeze in a quiet moment. Evenings are usually family TV time, so we need something easy to curl up on but still supportive. On weekends, we often have grandparents or friends visiting, so having a sofa that feels welcoming and can occasionally serve as an extra sleeping spot would be a bonus. More than anything, we'd like something that's comfortable, durable, and makes the room feel warm and lived-in.*
Best, Alex

Designer persona (30-min intake). Alex (she/her), 32; public school teacher; Queens, NYC (elevator, tight corridor turns). Home: 920 sq ft; living room ~ $14' \times 12'$; bay window + balcony door on one short wall. Household: partner + toddler + 35-lb beagle. Constraints: maintain $36''$ walkway; elevator delivery (modules $\leq 30''$; knock-down legs). Pain points: spills, pet claws, toy clutter; prior sofa too deep/sagged. Functional asks: modular, washable covers, storage ottoman, optional sleeper. Budget: mid. Consider: robot-vac clearance, circulation, modularity/delivery, kid/pet safety, easy cleaning.

Fig. 6. Version A living-room setting.

Version B.

Client email: Hi [your name], I'm looking for a sectional that makes my living room feel both comfortable and modern. The space is about 20′ × 16′ with a projector screen on one wall and low windows on the opposite side. There's also an opening into the dining area, so I need to keep a clear walkway along that edge. I often host friends for movie nights and board games, so the sofa needs to handle groups of 4–5 without anyone feeling cramped. Since I'm tall, I'd appreciate something that feels supportive for long stretches of sitting. I also work from home in sound design, so materials that feel durable and help the room's acoustics are important. I'd like something that fits the space without blocking the windows and keeps cables tidy around the projector. Comfort and longevity matter most; I don't need extra features beyond that. Best, Alex

Designer persona (30-min intake). Alex (he/him), 38; sound designer/post-production (WFH); Seattle, WA. Home: 2,200 sq ft; living room ~ 16′ × 20′ open-plan; projector on long wall; low-sill windows opposite; 9' ceiling. Household: lives alone; frequent movie/game nights. Pain points: echo/reverb; needs nearby power; dislikes compressing cushions. Ergonomics: supportive, taller back. Functional asks: large seating; breathable durable materials. Budget: upper-mid to premium (longevity/sustainability). Consider: seat geometry, low profile (avoid blocking windows), cable/power management, serviceability.

Fig. 7. Version B living-room setting.

B.2 Simulated Colleague (Stage 2 Handoff)

In Stage 2 (§4.3), participants received a standardized *simulated colleague* packet that mirrors the brief in Appendix B.1: three exemplar candidates plus a single-sentence designer note per candidate. The intent is to provide a consistent, asynchronous-by-proxy handoff against which participants compare their own favorites.

Table 1. Simulated colleague packets for both conditions. Each shows three candidate sofas (top row) and the one-sentence designer notes (bottom row).

Condition A (family condo; see §B.1)

		
Modular form works in small footprints while staying durable enough for kids and pets.	*The Chesterfield anchors the room with heritage character, giving the space a bold centerpiece.*	*Mid-century teal injects playful energy, brightening the room while keeping it family-friendly.*

Condition B (open-plan projector room; see §B.1)

		
Deep seats and warm neutrals soften the space, balancing comfort, acoustics, and sightlines.	*The tan leather Chesterfield adds timeless elegance that will age well and create cozy sophistication.*	*Bold magenta velvet transforms the room into a contemporary statement through color and texture.*

C Survey Instruments

C.1 Pre-study Survey

Name & Contact.

Q1. What is your full name (First, Last)? *[Open-ended]*

Q2. What is your email? *[Open-ended]*

Background.

Q3. How would you rate your English language proficiency? *[MC]*

- Native proficiency
- Advanced proficiency
- Intermediate proficiency
- Basic proficiency
- No proficiency

Q4. What is your gender? *[MC]*

- Female
- Male
- Non-binary/non-conforming
- Prefer not to respond

Q5. What is your age? *[Open-ended]*

Design Background.

Q6. What types of things have you designed before? (Defining the specific area helps us assess familiarity and experience.) *[Open-ended]*

Technical Background.

Q7. How often do you use Large Language Models (e.g., ChatGPT, DeepSeek, Gemini)? *[MC]*

- Never used
- Tried a few times
- Sometimes (a few times per month)
- Somewhat often (weekly)
- Regularly (daily)

Q8. How often do you use Image Generation models (e.g., Midjourney, DALL·E, Adobe Firefly)? *[MC]*

- Never used
- Tried a few times

- Sometimes (a few times per month)
- Somewhat often (weekly)
- Regularly (daily)

Q9. How often do you use Large Language or Image Generation models to assist with design-related work? *[MC]*

- Never used
- Tried a few times
- Sometimes (a few times per month)
- Somewhat often (weekly)
- Regularly (daily)

Confidence.

Q10. How confident are you right now in your designerly knowledge to design and visually illustrate a novel household item (e.g., chair, sofa, table)? *[1–7 Likert]*

- 1 — Not confident at all
- 2 — Very low confidence
- 3 — Somewhat low confidence
- 4 — Moderate confidence
- 5 — Fairly confident
- 6 — Very confident
- 7 — Completely confident

Q11. Please elaborate on why you rated your confidence at this level. *[Open-ended]*

Q12. How confident are you right now in your ability to use an Image Generation Model to design and visually illustrate a novel household item (e.g., chair, sofa, table)? *[1–7 Likert]*

- 1 — Not confident at all
- 2 — Very low confidence
- 3 — Somewhat low confidence
- 4 — Moderate confidence
- 5 — Fairly confident
- 6 — Very confident
- 7 — Completely confident

Additional.

Q13. Do you have any questions? *[Open-ended]*

C.2 In-study Survey #1 (Pre-creation Probe)

Learning.

Q1. List key *design dimensions* — the main axes you can vary to explore a sofa's design space (analogous to a table's height, surface material, edge shape). Please format each as a bullet beginning with "- ". *[Open-ended]*

C.3 In-study Survey #2 (Post–Stage 1)

Identification.

Q1. What is your Participant ID? *[Open-ended]*

Confidence.

Q2. Designerly confidence right now. *[1–7 Likert]*

- 1 — Not confident at all
- 2 — Very low confidence
- 3 — Somewhat low confidence
- 4 — Moderate confidence
- 5 — Fairly confident
- 6 — Very confident
- 7 — Completely confident

Q3. Did your designerly confidence shift from the pre-study survey? Why or why not? *[Open-ended]*

Q4. Confidence using an Image Generation Model right now. *[1–7 Likert]*

- 1 — Not confident at all
- 2 — Very low confidence
- 3 — Somewhat low confidence
- 4 — Moderate confidence
- 5 — Fairly confident
- 6 — Very confident
- 7 — Completely confident

Q5. Did your model-usage confidence shift from the pre-study survey? Why or why not? *[Open-ended]*

Stage-1 Reactions.

Q6. Right after this stage, what stood out to you? Did the range feel inspiring, overwhelming, or something else, and why? *[Open-ended]*

Q7. Rendering fidelity: would you prefer polished finished products, or to start with sketch-like/rough images? Why? *[Open-ended]*

Q8. With several AI options on screen, how did you decide what to refine or drop? *[Open-ended]*

Q9. For your last tweak, did you think in words, choose from tags, or something else? How did that fit your thinking? *[Open-ended]*

Q10. When exploring different directions, how did you create variety? What made it slow or effortless? *[Open-ended]*

Q11. How do you feel about starting sketch-like early and increasing detail as prompts get longer? *[Open-ended]*

C.4 In-study Survey #3 (Post–Stage 2)

Identification.

Q1. What is your Participant ID? *[Open-ended]*

Confidence.

Q2. Designerly confidence right now. *[1–7 Likert]*

- 1 — Not confident at all
- 2 — Very low confidence
- 3 — Somewhat low confidence
- 4 — Moderate confidence
- 5 — Fairly confident
- 6 — Very confident
- 7 — Completely confident

Q3. Did your designerly confidence shift from the 2nd in-study survey? Why or why not? *[Open-ended]*

Q4. Confidence using an Image Generation Model right now. *[1–7 Likert]*

- 1 — Not confident at all
- 2 — Very low confidence
- 3 — Somewhat low confidence
- 4 — Moderate confidence
- 5 — Fairly confident
- 6 — Very confident
- 7 — Completely confident

Q5. Did your model-usage confidence shift from the 2nd in-study survey? Why or why not? *[Open-ended]*

Collaboration & Comparison.

Q6. To what extent did you blend ideas from your own design and your partner's (simulated colleague's) designs? Describe your approach. *[Open-ended]*

Q7. When choosing between versions, what cues helped you notice key differences? What cues were missing? *[Open-ended]*

Exploration & Resumption.

Q8. How easy or difficult was it to explore ideas and track what changed or stayed the same? Why? *[Open-ended]*

Q9. Once you had a favorite, what made it easy or hard to branch without losing track of earlier ideas? *[Open-ended]*

Ownership.

Q10. When did the design feel most like "your" concept vs. an AI product, and why? *[Open-ended]*

Learning.

Q11. How did you learn about key design dimensions and possible values? *[Open-ended]*

Q12. Again list key *design dimensions* for a sofa (bullets starting with "- "). *[Open-ended]*

Enjoyment & Expressiveness.

Q13. How likely are you to use this system regularly, and why? (Typing or speaking aloud with auto-transcription is fine.) *[Open-ended]*

Q14. Did the AI produce something unexpectedly helpful or unhelpful? What led to that surprise? *[Open-ended]*

Q15. Did this tool help you better express your creativity? Why? *[Open-ended]*

Q16. Was the effort you put in worth it? Why? *[Open-ended]*

Feature Helpfulness.

Q17. Rank each feature by helpfulness (1 = least, 3 = most): *[Ranking]*

- Parameter-based deliberate variation (change specific tags within dimensions)
- Comparative interface for side-by-side inspection (aligned differences visible)
- Language support / terminology palette (dimension–tag palette scaffolding edits)

Q18. Why did you rank the features that way? *[Open-ended]*

Future Work.

Q19. What aspects of the design assistant were most intuitive or useful? *[Open-ended]*

Q20. What aspects were most confusing or difficult? *[Open-ended]*

Q21. Ideas to make exploring, comparing, and refining smoother — what would that look like? *[Open-ended]*

D Needs–Interventions Mapping

This appendix elaborates the literature-derived needs that underlie the synthesis in §2.4. We separate *pre-GenAI* needs (before text-to-image/LLM tools) from *post-GenAI* needs (when tens–hundreds of alternatives become routine). "Mesoscale" denotes dozen-level candidate sets accumulated across iterative turns, where light structure helps track differences and rationale.

Table 2. Needs and primary interface levers. Levers: *Variation* (controlled changes to factors), *Comparison* (aligned side-by-side views), *Terminology* (name and stabilize factors), *Fidelity* (stage sketch↔render). References are representative.

Need (abbr.)	Asynchronous pain point	Primary intervention(s)	Representative refs
Pre-GenAI (N1–N4)			
N1 Coordination/Awareness	Who changed what and why is opaque; weak provenance across turns	Terminology, Variation, Comparison	[46, 84]
N2 Structured Comparability	Unaligned alternatives; tab overload; trade-offs not explicit	Comparison	[13, 59, 60]
N3 Between-session Reflection	Momentum and rationale dissipate between turns	Terminology, Variation, Comparison	[33, 56, 82]
N4 Externalized Rationale	Intent and criteria remain implicit; poor handoff quality	Terminology	[37, 79, 85]
Post-GenAI (N5–N8)			
N5 Mesoscale Sensemaking*	Candidate overload; subtle differences missed in large batches	Variation, Comparison, Fidelity	[7, 26, 48, 74]
N6 Convergence with Diversity	Premature homogenization; loss of exploratory coverage	Variation, Comparison, Fidelity	[20, 21, 23]
N7 Teachable Dimensions	Tacit criteria unnamed; terms drift across collaborators	Terminology	[19, 41, 69, 74, 82]
N8 Shareable, Replayable Strategy	Exploration strategy cannot be transferred or reproduced later	Terminology, Variation, Comparison	[17, 18, 64, 81]

* *Mesoscale* = dozen-level candidate sets across iterative turns.

Table 3. How needs (N1–N8) map to design goals (DG1–DG3). DG1: dimensions and terminology; DG2: intentional variation and aligned comparison; DG3: staged fidelity.

Need	DG1	DG2	DG3
N1 Coordination/Awareness	✓	✓	
N2 Structured Comparability		✓	
N3 Between-session Reflection	✓	✓	
N4 Externalized Rationale	✓		
N5 Mesoscale Sensemaking		✓	✓
N6 Convergence with Diversity		✓	✓
N7 Teachable Dimensions	✓		
N8 Shareable, Replayable Strategy	✓	✓	

Interpretation. DG1 (dimensions and terminology) externalizes criteria and stabilizes language across turns and collaborators (N4, N7), improving awareness and transfer of strategy (N1, N8) and sustaining reflection (N3). DG2 (intentional variation and aligned comparison) makes alternatives commensurate and trade-offs explicit (N2), supports mesoscale sensemaking (N5), and enables principled convergence without collapsing diversity (N6), while also helping maintain

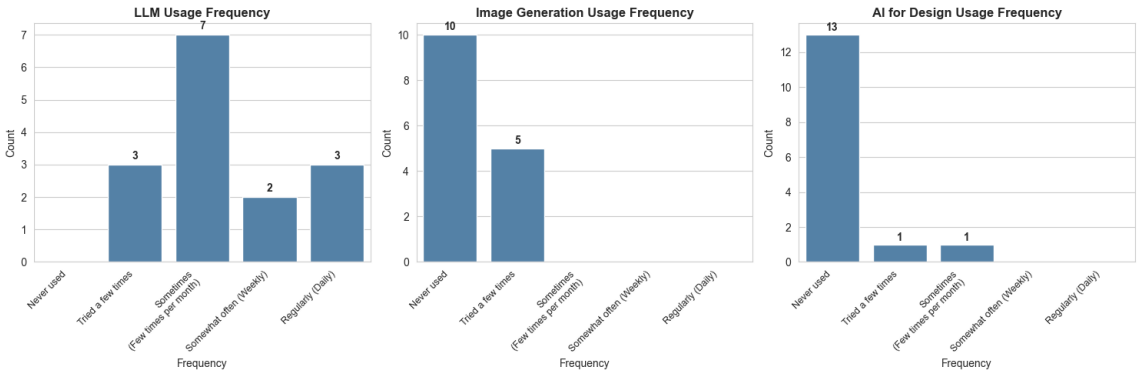

Fig. 8. Self-reported prior exposure to AI tools before the study (N=15). Panels show frequency categories for three tool types: large-language-model (LLM) use, text-to-image use, and "AI for design." Values are counts.

momentum (N3, N8). DG3 (staged fidelity) aids perception and evaluation at scale (N5) and complements convergence while preserving breadth early (N6).

E Supplementary Descriptives and Plots

Participant background (descriptive). Before the study, LLM use ranged from infrequent to daily (most reported using them monthly or less); most had *never* used text-to-image systems or "AI for design." For context only (no inferential tests): LLM usage counts were *3 tried a few times, 7 sometimes (monthly), 2 weekly, 3 daily*; image generation *10 never, 5 tried a few times*; AI-for-design *13 never, 1 tried a few times, 1 sometimes*. These distributions are provided to contextualize qualitative findings in §4; no between-group comparisons are made.

Received September 11, 2025