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

ANONYMOUS AUTHOR(S)
SUBMISSION ID: 8966

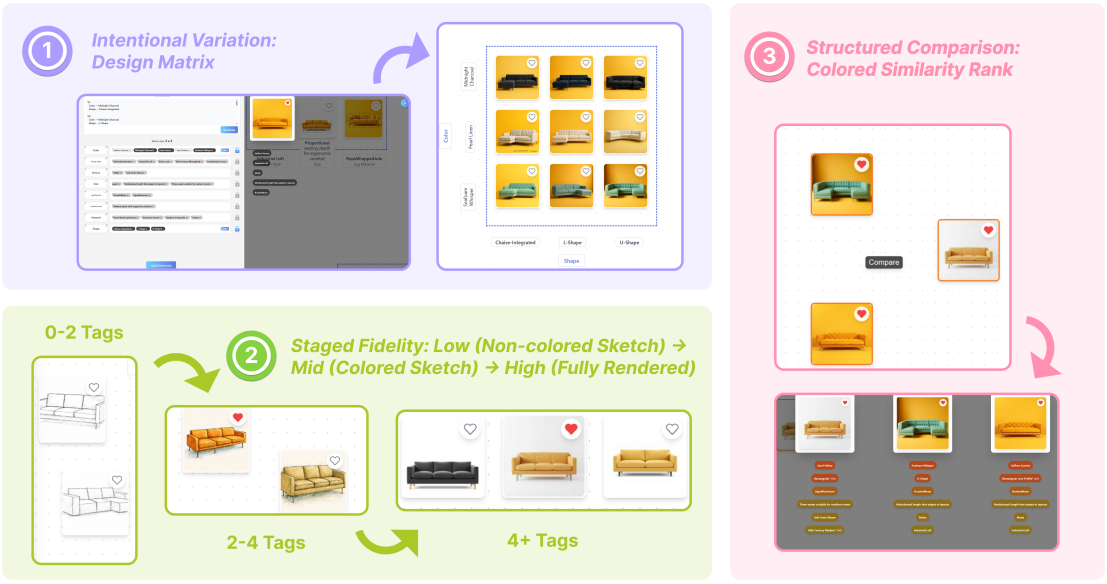

Fig. 1. Main features for *VariationWeaver*: (1) intentional variation using a attribute-based design matrix; (2) staged fidelity from sketch to colored to fully rendered; and (3) structured comparison with colored similarity rank.

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. We examine how coordinating three mechanisms—intentional variation, staged fidelity, and structured comparison—shapes how novices explore and converge with a text-to-image model. We present *VariationWeaver*, a canvas interface that instantiates these mechanisms over a tagged dimensional palette for novice product ideation. In a qualitative lab study with 15 novice designers, we found that: 1) intentional variation helped participants notice and name key design dimensions while keeping changes incremental and controllable; 2) structured comparison supported trade-off reasoning when variants differed along clearly labeled dimensions; and 3) sketch-first outputs sometimes provoked new ideas yet often felt too ambiguous for final evaluation. Together, these patterns suggest concrete design implications for how GenAI tools coordinate variation, comparison, and fidelity, and they also raise tensions for simple "low-fi first" assumptions once detailed renders are readily available.

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, 40 pages. https://doi.org/10.1145/XXXXXXX.XXXXXXX

1 INTRODUCTION

Advances in text-to-image (T2I) generation enable amateur designers create photo-realistic images of product ideas from short natural language prompts [80, 83]. While the high-fidelity outputs of modern T2Is can be impressive, the theory on design and prototyping would suggest that polishing visuals too early can induce fixation, draw attention to surface features, and discourage lateral moves in the design space [11, 41, 85, 102]. Sketching is typically seen as a methodological antidote to converging prematurely because they are not only fast to make, but also because they leave room for reflection in action and external critique [87, 99]. Information foraging theory similarly suggests that people follow cues with strong "information scent" when navigating complex spaces [78]; when GenAI tools emphasize a single evolving prompt and a few striking high-fidelity samples, they may foreground visually salient cues at the expense of exploring the underlying design dimensions. For all the rhetoric on how GenAI will disrupt creativity, current tools do a poor job of supporting effective designerly practices like ramping up fidelity [65, 102] and creating parallel prototypes [21, 22, 33].

To unpack this tension, we draw on work on variation and prototype fidelity. Variation theory argues that people discern critical features when one aspect of a situation varies while others remain stable, turning near-miss contrasts into learning signals [33, 63]. At the same time, sketching and prototyping work highlights the value of sweeping broadly across a space with low-fidelity alternatives to see qualitatively different possibilities [11, 102]. Together with prior parallel prototyping findings introduced above, this suggests a productive tension: early in a project, designers benefit from broad moves that traverse diverse regions of the space, whereas later they need interfaces that support narrower, one-factor refinements around salient dimensions. Work on low- versus high-fidelity artifacts reinforces this: sketch-like, ambiguous representations invite speculation and reinterpretation, while polished renderings support precise evaluation but can prematurely narrow the space [14, 26, 96]. Effective GenAI tools should therefore support both expansive early exploration and more focused, one- or two-factor refinement as concepts converge, while staging fidelity to match designers' evolving intent.

Recent HCI systems begin to encode parts of this agenda in GenAI-based design tools. DesignWeaver, for example, surfaces domain-relevant language and palettes of prompt fragments to help novices articulate T2I product briefs [97]. Other systems structure how people express and organize prompts—using LLM-enhanced templates on whiteboards, semantic labels on mood boards, and visual programming for LLM chains [48, 104, 106]—scaffold exploration of many alternatives at once through iterative and layout-based generation interfaces [2, 18], and explicitly organize design responses along labeled axes and keywords to expose a structured design space [16, 94]. However, despite these advances, current systems still leave important aspects of structured exploration underdeveloped. As a result, we identify three gaps in how current tools support structured exploration of GenAI-based product design spaces: first, the design space often remains *tacit*, with alternatives generated without clearly naming the underlying dimensions or terminology [52];

second, while interfaces may show multiple images at once, differences are rarely presented in *commensurate*, aligned formats or explicitly labeled, so users must rely on superficial visual inspection rather than structured comparison [27, 77]; third, existing tools seldom connect variation to explicit control over prototype fidelity and ambiguity, despite longstanding evidence that low- and high-fidelity artifacts support different kinds of reasoning and critique [14, 26, 96].

We introduce *VariationWeaver*, an interactive canvas for T2I product ideation that builds on DesignWeaver's palette-based language scaffolding [97], but adds control over how alternatives vary, how fidelity ramps, and how outputs are compared. First, *intentional variation* lets designers pick one or two palette dimensions to vary in each batch while holding the rest fixed, showing alternatives with tagged dimensions so novices can acclimate to the design space and see how specific changes reshape designs. Second, *staged fidelity* starts with sketch-like images and increases polish only as prompts become more specific, so high-fidelity renders appear after designers commit to concrete constraints. Third, *structured comparison* ranks images by model-encoded semantic distance to surface the most distinct alternatives, and uses gradient color coding over tag differences to surface key distinctions.

We evaluate *VariationWeaver* in a qualitative lab study with 15 novice designers working on a sofa design brief. The study focuses on how the canvas shapes exploration, refinement, and the experience of using T2I tools for product design, asking:

(1) **RQ1: Features and design trajectory.** How do *VariationWeaver*'s three mechanisms—intentional variation, structured comparison, and staged fidelity—shape how novices explore, refine, and converge on T2I-based sofa designs?
(2) **RQ2: Experience.** How do novices feel overall about using *VariationWeaver* for product design?

Across sessions, novices used intentional-variation controls to acclimate to unfamiliar dimensions, learn vocabulary, and adjust sofas by changing few factors at a time. Structured comparison helped some participants read specific differences and justify choices, but felt redundant or confusing when alternatives were too similar or labels were unclear. The sketch-first policy encouraged a subset to treat early images as rough ideas to improve, while others felt unable to evaluate sofas until polished renders appeared. Confidence and ownership tended to increase when targeted edits produced visible, incremental changes aligned with participants' intentions or when they could borrow and adapt colleagues' designs; when the model ignored targeted edits or outputs stayed far from their ideas, participants reported lower agency and weaker ownership.

This paper contributes:

- **System.** *VariationWeaver*—a working canvas for T2I product ideation that operationalizes intentional variation scaffolded with dimensional language, staged fidelity policies, and structured comparison views.
- **Empirical insights.** A qualitative account of how these features shape novices' design trajectories: how intentional variation support dimensional acclimation and controllable change; how staged fidelity and structured comparison can both help and hinder noticing differences and making trade-offs; and how these dynamics influence confidence, perceived control, and ownership. We distill these observations into design guidance for future generative ideation tools.

2 RELATED WORK

We extend the introduction's arguments about variation, fidelity, and comparison by connecting *VariationWeaver* to three strands of prior work: (1) how designers explore and structure variation in high-dimensional spaces; (2)

how fidelity and ambiguity shape ideation when photorealistic renderings are fast to obtain; and (3) how external representations align alternatives for comparison and coordination.

2.1 Exploring and Structuring Variation in High-Dimensional Design Spaces

Design work is often framed as navigating high-dimensional spaces of alternatives, where progress depends on surfacing, naming, and reusing salient dimensions rather than treating each artifact in isolation [5, 20, 52, 88]. Instruction-tuned language models can transform informal prompts into structured briefs and domain descriptions [10, 17, 69], while modern text-to-image pipelines support flexible image generation and editing [12, 81, 83], with extensions to 3D assets [53, 55, 79] and to video or interactive media [3, 4, 37, 49, 91]. HCI systems wrap these models to structure how people express and organize variation: some provide prompt templates [24, 106], semantic labels on mood boards [48], or whiteboard–LLM workflows that guide prompt construction and envisioning [56, 93, 105]. Others support iterative change through sketch-based tuning [15], region- and story-based editing [18], accessible refinement for blind and low-vision creators [39], or block- and voxel-based workflows [84, 90]. Interfaces for browsing large sets of outputs use parameter sweeps [61], clustered galleries [7], and graph-structured prompt–image spaces [42] to help users scan and cluster alternatives. Recent work further treats generative outputs as explicit *variation spaces*: exposing navigable axes and tags [30, 94, 95], treating style and visual properties as editable parameters [44, 45], or automatically expanding prompts to increase diversity and personalization [32]. Spreadsheet-like tools let designers systematically manipulate prompts and recombine references [2, 16, 92], while gallery and history views help people make sense of many outputs over time in text-to-image workflows [28, 97]. Similar abstraction and aggregation techniques organize large variation spaces in programming and generative design, using clustering, exemplar selection, and constraint reasoning to reveal structure [5, 29, 33, 64, 74, 103]. Parallel prototyping studies show that aligned "near-miss" sets of alternatives can improve critique quality and downstream decisions [21, 22, 98], and ProcessGallery further demonstrates how contrasting early and late iterations helps learners connect controlled changes to design principles [107].

Collectively, this work shows how generative models can be wrapped in parametric and gallery-based interfaces that *expose* many dimensions. DesignWeaver, in particular, stabilizes domain-relevant language by attaching tags and palette axes to images, making key dimensions more explicit for novice designers [97]. However, most systems present variation as loosely structured grids or canvases: novices see many alternatives but have limited support to *intentionally construct* aligned sets where only one or two palette dimensions change while others remain stable, and variation is rarely coordinated with explicit policies over output fidelity. *VariationWeaver* extends this line of work by coupling palette-based dimensional scaffolding with intentional-variation controls and staged-fidelity policies that generate compact, aligned sets of one- or two-factor changes across designs.

2.2 Fidelity and Ambiguity When Photorealistic Renderings Are Fast to Obtain

Classic prototyping work argues that low-fidelity artifacts keep ideas inexpensive to change and make critique safer, whereas higher-fidelity representations support specification and detailed evaluation [11, 85, 102]. Early exposure to polished forms can induce fixation on surface features and hinder exploration [41]. Reviews and protocol studies show that sketches enable reinterpretation and discovery of new spatial or functional relations beyond the designer's initial intent [43, 96], and that ambiguity itself can be leveraged as a resource for interpretation and reflection [26]. Empirical work documents how low-fidelity representations support early ideation even as organizational and cultural pressures push practitioners toward higher-fidelity prototypes [14, 67, 108]. These accounts collectively motivate keeping early artifacts rough to encourage broad exploration and critical feedback before committing to detail.

Systems research codifies some of these practices into structured canvases that surface domain-relevant dimensions and vocabulary for novices [20, 52, 88]. For instance, ImaginationVellum integrates spatial prompts, generative strokes, and history to support ongoing exploration while preserving sketch-like cues [62]. However, the role of fidelity changes when text-to-image pipelines make photorealistic renderings comparatively fast and inexpensive relative to manual sketching or 3D modeling [14]. It is less clear whether "sketch first, render later" remains the only productive strategy, or how fidelity should be staged as designers' intent evolves. In this work, we embed a staged-fidelity policy into *VariationWeaver*—delaying fully polished images until users specify more detailed constraints—and use the resulting interactions as a probe into how GenAI capabilities interact with prior claims about fidelity, ambiguity, and exploration.

2.3 External Representations that Align Alternatives for Comparison and Coordination

Work on external representations shows that the structure of artifacts outside the head shapes how people think, remember, and coordinate [1, 86, 110]. Aligned, side-by-side layouts foreground structural matches and make correspondences easier to see [27], while joint evaluation studies show that viewing options together changes how people trade off attributes and can reduce cognitive load when alternatives are presented in normalized grids [38, 76, 77]. Gallery-style canvases and shared displays help teams notice and name differences, remix and fork ideas, and converge on commensurate representations for discussion [13, 34, 109]. Choice-support systems and dynamic labeling interfaces further demonstrate how structured displays and evolving labels can guide reflection and multi-stakeholder decision making [72, 73]. In GenAI settings, variation interfaces such as Luminate and related gallery tools organize large sets of generated options along explicit axes or clusters, helping users scan and compare many designs at once [28, 94]. Recent work also suggests that text-to-image artifacts can catalyze shared semantics without fully collapsing the design space [54], while strategy-centric guidance and templated prompting can homogenize ideas and weaken later originality [50].

Collaborative design work depends on maintaining shared frames; misaligned frames create friction and rework, especially in early phases [23, 36, 47]. Studies of sensemaking artifacts and engineering handoffs show that artifact structure and completeness strongly affect how easily later contributors can resume and extend work [75, 89]. Distributed and crowd-based workflows rely on standardized representations, preserved decision trails, and clear task decomposition to coordinate contributions across time and roles [46, 68, 82]. Awareness mechanisms and interruption studies highlight the importance of cues that support resumption and between-session reflection [31, 40], while rationale schemes and visualizations of activity traces make disagreement and multi-criteria decisions more inspectable [57, 60, 101]. Recent tools for logging prompts and mixed-initiative workflows begin to capture GenAI-era design traces [19, 104], and interaction theory frames canvases as *interaction substrates* that organize data, constraints, and tools so users can manipulate objects of interest in principled ways [59]. *VariationWeaver* is inspired by this view: by making dimensions explicit, organizing alternatives through intentional variation and structured comparison, and staging fidelity over time, we investigate how a T2I canvas can better support comparison and within-session coordination while laying groundwork for more robust asynchronous use.

3 TOOL DESIGN: Variation, Fidelity, and Comparison

3.1 Design Goals

Building on DesignWeaver's palette-based language scaffolding as a starting point [97], *VariationWeaver* adds three design goals that organize how variation, comparison, and fidelity work together. These goals respond to recurring needs we synthesize from prior work and detail in Appendix D: tools should (1) help novices notice, name, and reuse

salient dimensions; (2) support structured, intentional variation and commensurate comparison so alternatives remain diverse yet comparable; and (3) stage fidelity so early exploration stays sketch-like and low-commitment, with detail increasing as intent stabilizes.

DG1 — Intentional variation over tagged dimensions. Provide controls that let designers vary one or two tagged dimensions at a time while holding others fixed, using a stable palette of terms and exemplars so novices can acclimate to key design dimensions and see how specific changes reshape sofas without losing prior context.

DG2 — Structured, commensurate comparison. Arrange alternatives in compact, aligned views with shared tags, ranking images by model-encoded semantic distance and tag differences to foreground informative contrasts so differences—and, when present, trade-offs across designs—are easier to read.

DG3 — Staged fidelity. Begin with sketch-like outputs to support broad, low-commitment exploration and reinterpretation, then increase fidelity as prompts become more specific so detailed evaluation occurs once concepts have stabilized and constraints are clearer.

3.2 User Experience

VariationWeaver is designed for short, iterative sessions where novices move from a vague brief to a small set of curated options. Figure 2 illustrates the intended workflow through "Maya," a novice designer responding to a compact-sofa brief. *VariationWeaver* structures her work into eight steps that make dimensions explicit, encourage controlled variation, and capture rationale for later review.

Maya first reviews the design material and types a plain-language prompt in the sidebar (Brief, Prompt). *VariationWeaver* parses the brief, suggests initial dimensions such as era, color, material, and shape, and seeds each row of the palette with a few illustrative tags (Build Palette). These tags give Maya concrete handles for talking about the design space rather than repeatedly editing raw text prompts.

She then alternates between generating structured variation and judging results. For each move, she chooses up to two dimensions to vary while freezing the rest—for instance, sweeping color and leg style while holding shape, size, and upholstery constant. *VariationWeaver* returns near-miss sets in aligned grids on the canvas so that differences between variants are easy to read (Vary, Compare). Maya zooms in on promising clusters, edits or adds tags, and re-runs variation to explore targeted "what-if" changes (Refine).

Once she is satisfied, Maya marks a small set of favorites, writes short notes about why specific combinations work, and compiles a shortlist for the client (Curate). In an asynchronous handoff, a colleague opens the same canvas, sees Maya's favorites and comments in place, and can either choose among them or launch new variation runs from the saved palette (Asynchronous Handoff). The journey emphasizes that *VariationWeaver* is not only a generator, but a shared canvas for evolving and communicating a design space across iterations.

3.3 Implementation Details

VariationWeaver combines a prompt-and-palette sidebar with an infinite image canvas that organizes generated sofas into structured groups (Figure 3). The sidebar stabilizes terminology and intent; the canvas visualizes variation and comparison; a compact toolbar supports lightweight comments and favorites used in the asynchronous phase.

Prompt box and dimension palette (Component A). Designers begin in the Design Sidebar by reading the brief and entering an initial prompt. Below the prompt box, the Dimension Palette lists each active dimension as a row (for example, Color, Design Style, Texture, Size, Material, Shape) with chip-like tags for specific values (Figure 4). Users can add new dimensions, rename or delete existing ones, and edit tags in place. A tag-edit dialog shows example images

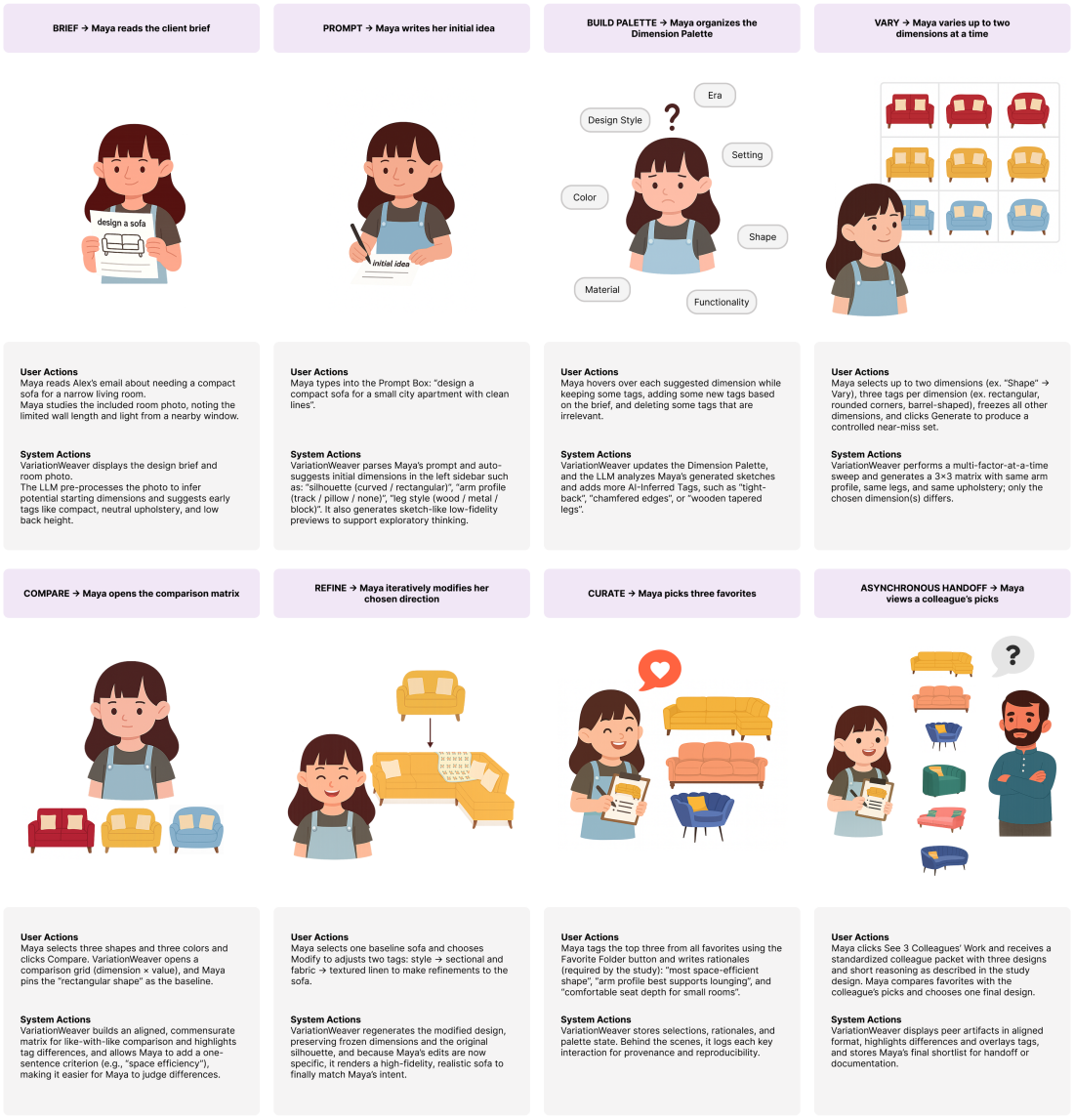

Fig. 2. End-to-end user journey in *VariationWeaver*. Maya moves from a brief and initial prompt to building a dimension palette, generating structured variation, comparing options, refining, curating favorites, and asynchronously handing off her canvas to a colleague.

and short explanations (for instance, "Saffron Sunrise: warm, golden-orange fabric that adds sunny vibrancy to sofas"), helping novices develop a concrete sense of what each term means. Because tags are reused across runs and attached to generated images, the palette acts as a stable vocabulary that grows with the project.

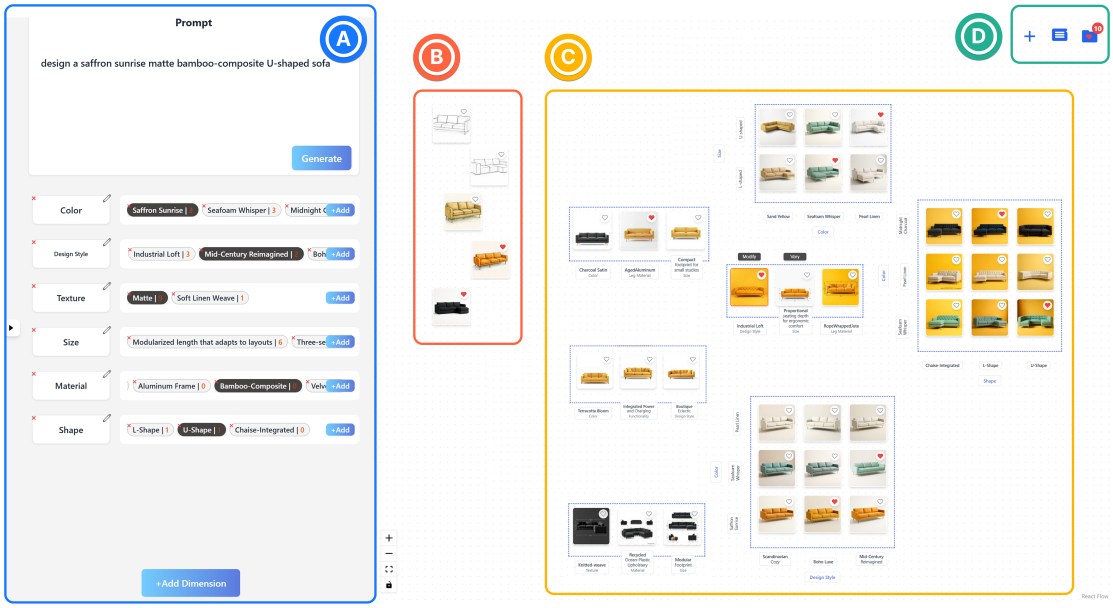

Fig. 3. *VariationWeaver* interface overview for the sofa design task. (A) Prompt box and dimension palette where users compose the brief and manage tagged dimensions; (B) column showing a fidelity progression of the currently highlighted sofa set; (C) main canvas with aligned grids of generated sofas organized by varied dimension; and (D) global controls for sidebar, comments, and favorites.

Image canvas for structured variation and comparison (Components B and C). The right side of the interface is a zoomable canvas where generated sofas are laid out in small grids (Figure 5). Each grid corresponds to a variation batch in which one or two dimensions change while others are held constant. A vertical strip of thumbnails shows a fidelity progression for the current batch. Clicking an image opens an inspection panel with AI-inferred tags, making it easy to see which palette values the model actually realized. From any inspected image, users can trigger *Modify* (regenerate within the same tag combination) or *Vary* (sweep one or two dimensions while freezing the rest). A simple color-rank view lines up candidates along a single dimension, helping users check coverage and gaps in the palette.

Lightweight comments and favorites (Component D). To support asynchronous review, designers can add comments directly on the canvas, like images, and mark a small set as favorites (Figure 6). The toolbar at the top-right toggles the sidebar, creates comment pins, and opens a favorites-only view used later in the simulated-colleague phase of the study. We intentionally kept these collaboration features simple so that the primary interaction remains exploring and comparing aligned variation sets.

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), implementing the staged-fidelity policy. During comparison, an axis-ordering heuristic surfaces dimensions with greater dispersion within the current set to highlight informative contrasts.

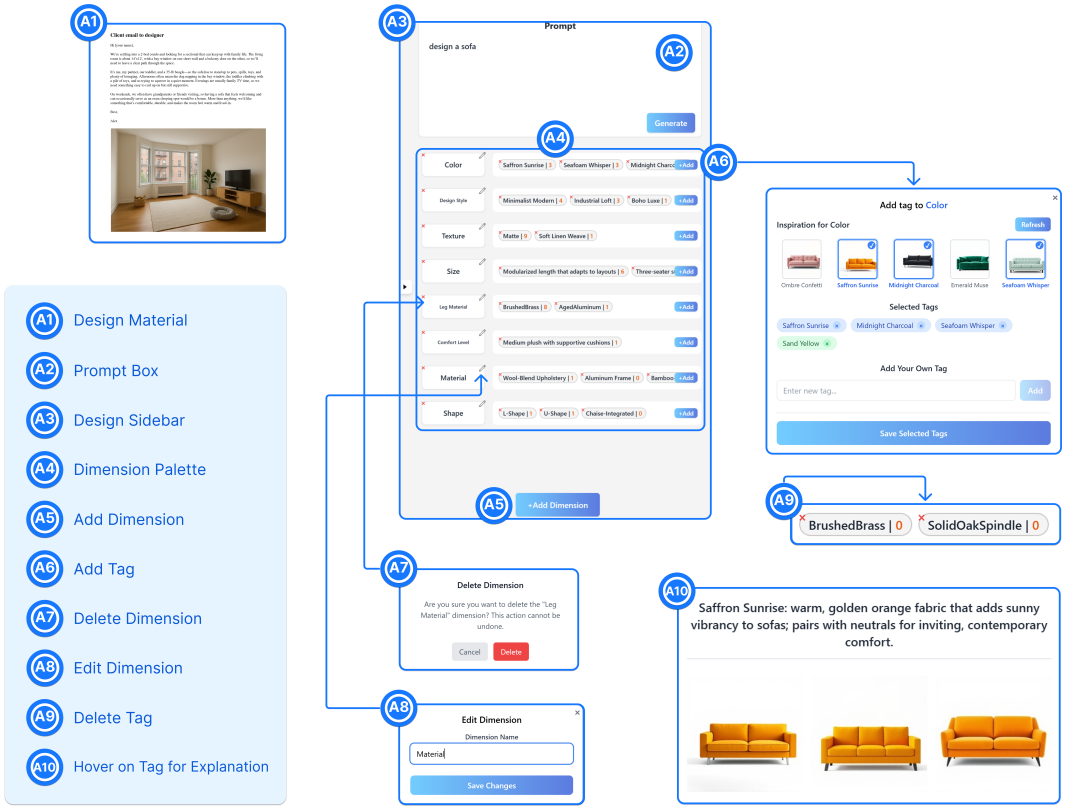

Fig. 4. Component A: Design Sidebar. Designers read the brief, enter a prompt, and build a dimension palette by adding, editing, or deleting dimensions and tags with inline controls and explanatory examples.

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 are available in an anonymized repository at https://github.com/mushroom-labs/variation-weaver-chi26.

3.4 Generalizability

VariationWeaver is not tied to sofas; its palette, intentional-variation controls, and comparison canvas are domain-agnostic. Figure 7 and Figure 8 show lamp and toy-packaging briefs with different dimensions (for example, shape, material, functionality, and sustainability). We chose sofas for the study because they are familiar and easy to critique, but the same workflow extends to other everyday products with only prompt and palette changes.

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.

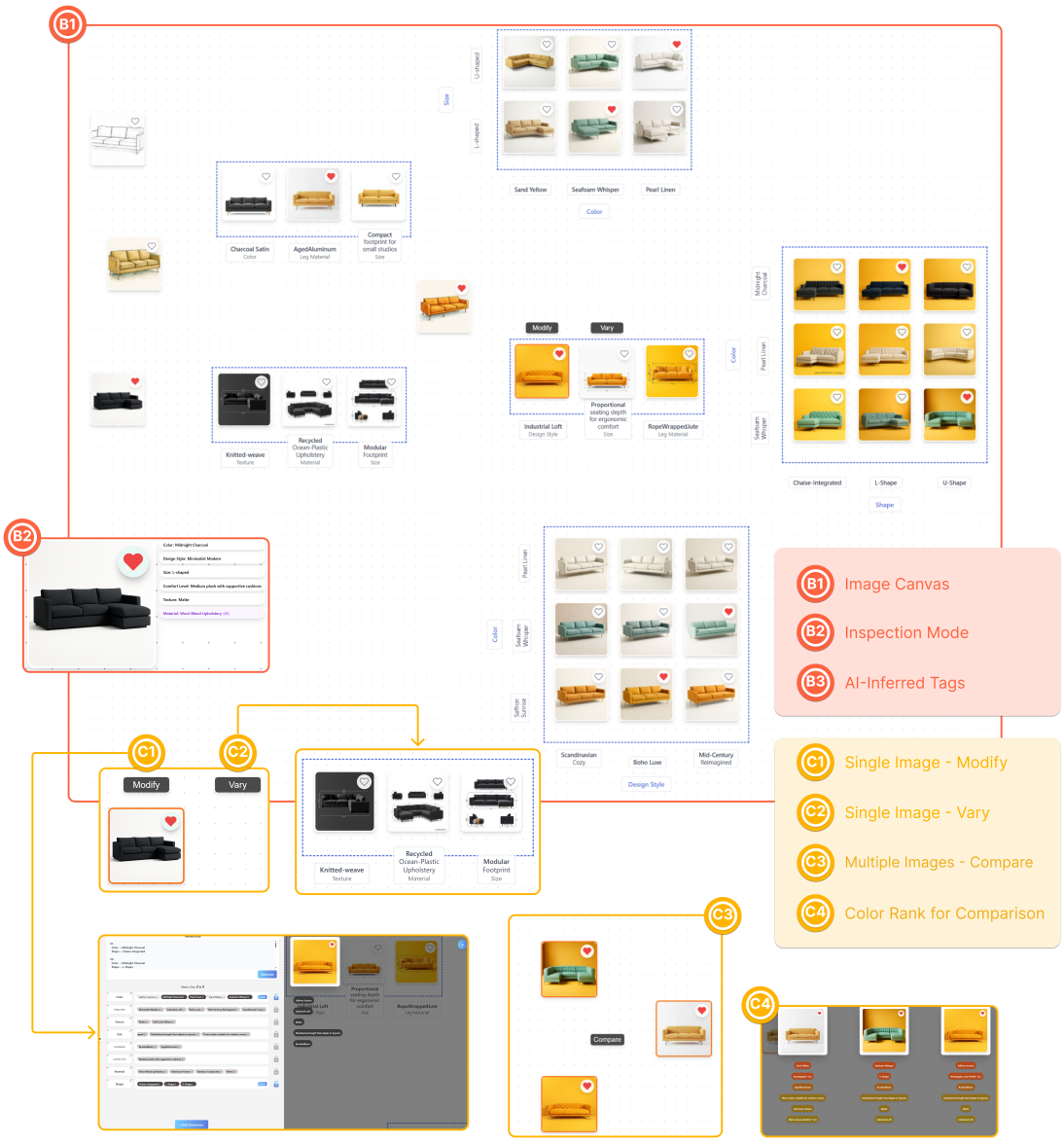

Fig. 5. Components B and C: image canvas and variation controls. Grids on the canvas show batches where one or two dimensions vary; inspection mode reveals AI-inferred tags, and Modify / Vary / comparison views support controlled near-miss sets and side-by-side inspection.

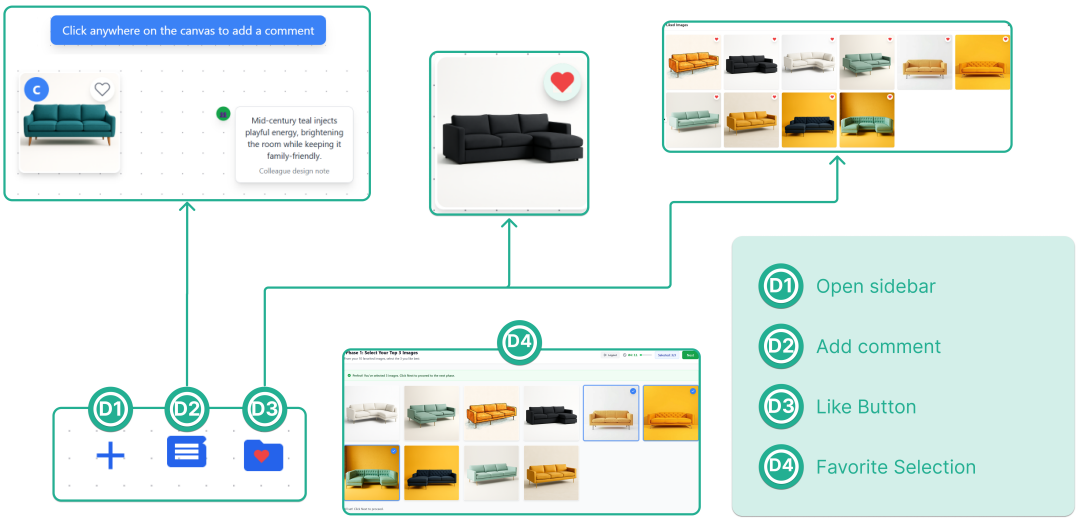

Fig. 6. Component D: comments and favorites. Designers can pin comments on sofas, like images, and mark a subset as favorites; toolbar buttons open the sidebar, add comment pins, and jump to the favorites view.

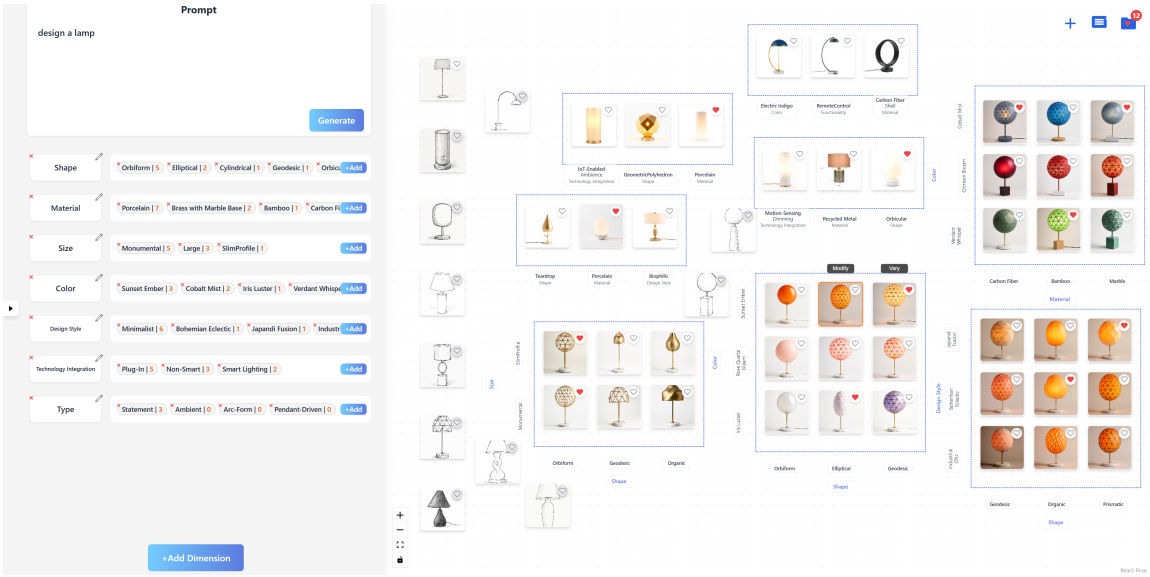

Fig. 7. *VariationWeaver* used for lamp ideation with lamp-specific dimensions and aligned grids of generated variants.

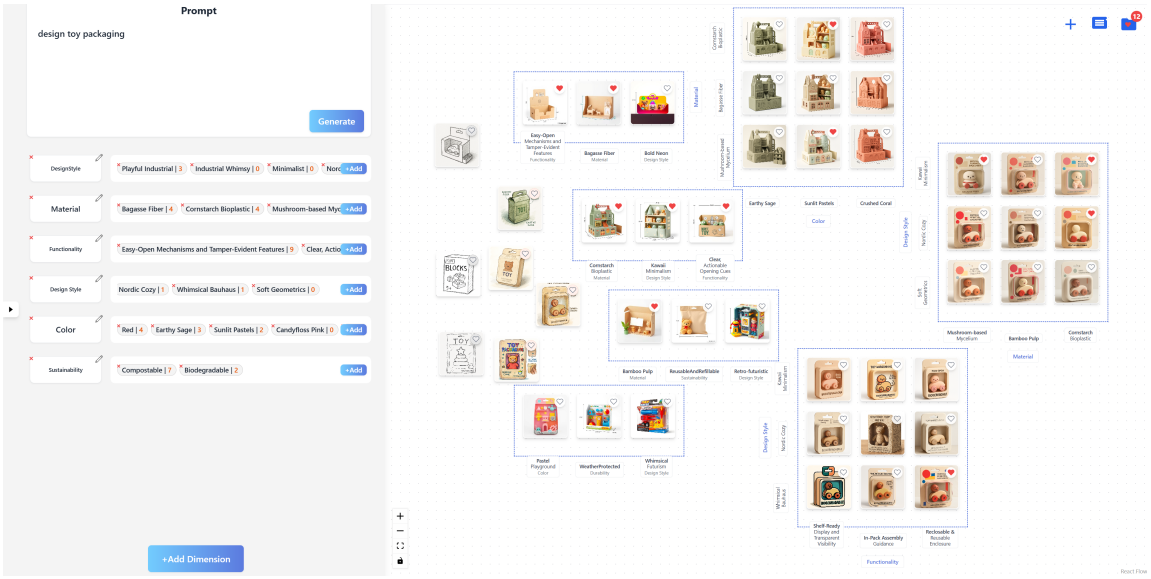

Fig. 8. *VariationWeaver* used for toy-packaging ideation with packaging dimensions and aligned grids of variants.

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

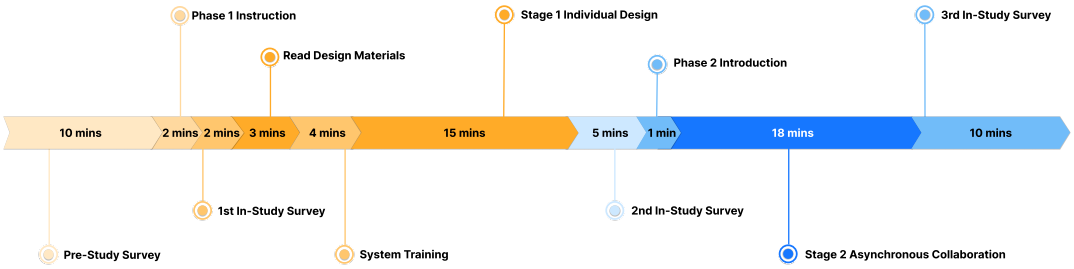

Fig. 9. Study session timeline: pre-survey, tutorial, Stage 1 (individual exploration; top-3 + rationales), probe, Stage 2 (simulated handoff; compare; shortlist + rationales), post-survey and interview.

is summarized descriptively in Appendix E (Fig. 14): 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; 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 9.

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]. The process included 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 [35, 66]. Critical-friend peer debriefs [100] followed HCI-methods guidance [6, 51]. 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. To ground the discussion, Figure 10 shows Participant P2's staged progression from a sketch-like preview to higher-fidelity variants across Stage 1 and Stage 2. Findings are organized by our research questions: **RQ1** on how *VariationWeaver*'s mechanisms for intentional variation, structured comparison, and staged fidelity shaped novices' design trajectories, and **RQ2** on how novices experienced the tool (for example, confidence and ownership). A ranking probe provides descriptive context (Figure 11); these counts are not inferential.

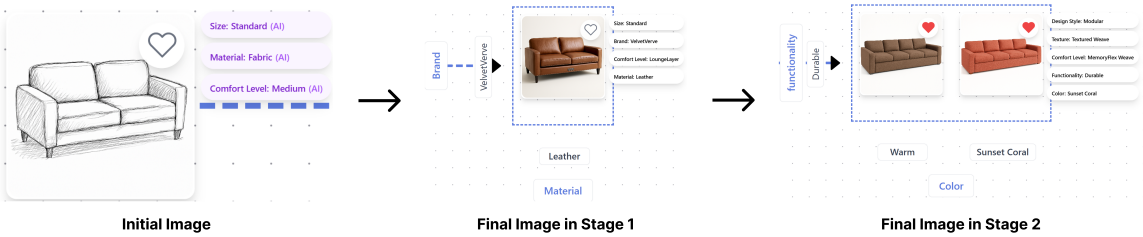

Fig. 10. P2's staged progression from an initial sketch-like sofa to higher-fidelity variants across Stage 1 and Stage 2.

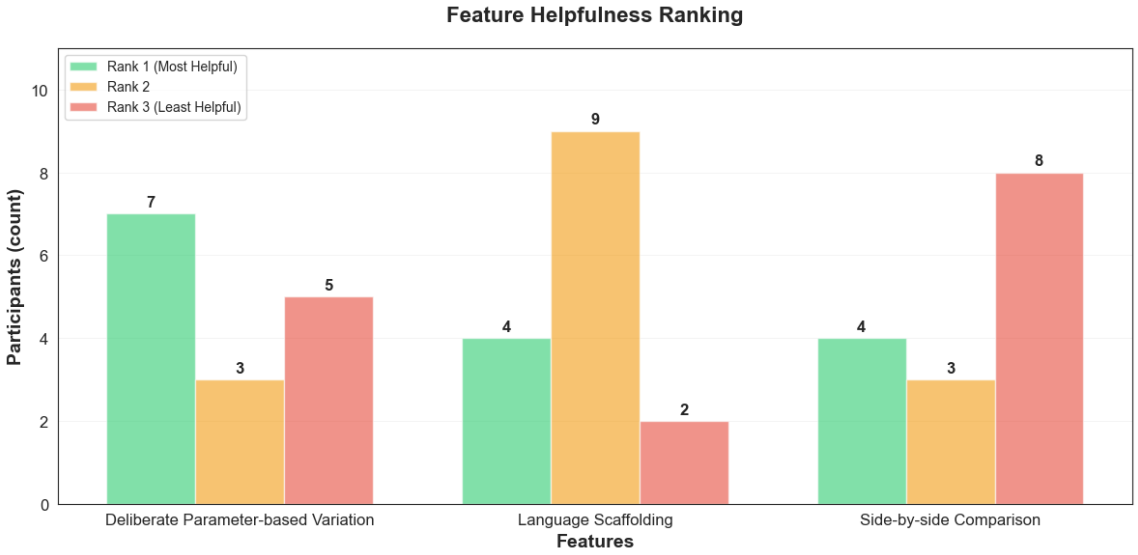

Fig. 11. Self-reported helpfulness ranking (N=15). Variation received the most Rank-1 votes; language scaffolding most Rank-2; comparison most Rank-3.

5.1 RQ1: Features and design trajectory

RQ1 asks how *VariationWeaver*'s three mechanisms—intentional variation over tagged dimensions, structured comparison, and staged fidelity—shaped how novices explored, refined, and converged on sofa designs.

[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 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*").

[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. 11).

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).

5.2 RQ2: Experience—confidence and ownership

RQ2 examines how novices felt about designing with *VariationWeaver*, focusing on shifts in confidence and perceived ownership of the resulting sofas.

[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 through the lens of our three mechanisms—intentional variation (implemented via the palette), structured comparison, and staged fidelity—and explain how they produced both benefits and friction for novices. We then connect these observations to prior work and outline focused directions for future systems and studies.

6.1 Intentional variation via the palette as acclimation scaffold

Participants acclimated fastest when they could both *see* key dimensions and *vary* them deliberately using tagged controls (Theme 1). Rather than learning vocabulary in isolation, they learned by doing: selecting tags, freezing others, and observing how a controlled change reshaped a set of sofas. This pattern is consistent with Information Foraging Theory: strong "information scent"—clear labels with immediate payoffs—reduces navigation cost and supports learning [70, 71, 78]. It also aligns with Variation Theory, which emphasizes that systematic contrast across examples helps people discern what matters [63]. The palette's chips, hover previews, and tag suggestions act as external representations that offload cognition and help people map terminology onto perceptual differences [1, 58, 86].

Data from this study suggest that intentional variation is the mechanism that turns the palette from a static glossary into an acclimation scaffold. When participants edited tags or added new ones, they were effectively specifying which dimensions should change and which should stay fixed, then using image sets to inspect the outcome. The same visible lexicon that teaches words thus defines the axes along which users can systematically explore. When labels were missing or ambiguous—or when the model did not respond to tagged changes—this structure broke down and participants fell back to vague prompting, reporting higher search cost and weaker control.

Design implication. Design palettes should be treated as both a durable, learnable lexicon and a control surface for intentional variation. Interfaces should pair each term with a small exemplar, keep terms stable across turns, and

support image→tag remapping so users can pull language from promising images rather than inventing all descriptors upfront.

6.2 Structured comparison depends on meaningful variation

Structured comparison helped when alternatives were commensurate and differences were diagnostic (Theme 2). Structural alignment theory predicts that aligned formats make correspondences legible and support relational comparison [27]. Joint, side-by-side evaluation also reduces memory load and helps people articulate trade-offs [38, 76, 77]. Positive accounts from P2, P4, P10, and P15 match this pattern: when axes were clear, and when tag differences reflected visible changes, the matrix made it easier to see what each choice gained and lost.

By contrast, comparison faltered when sets were too homogeneous, when labels were opaque, or when the model ignored targeted edits. Variation Theory argues that without systematic contrast—varying one or a few dimensions while others remain stable—learners struggle to discern what matters [63]. In this study, tightly clustered images with unclear differences reduced comparison to "spot the trivial change," and model non-compliance collapsed the intended structure altogether, pushing people back to guessing or visual scanning alone. Participants then described the interface as redundant or confusing, even though the layout itself had not changed.

Design implication. GenAI interfaces should coordinate comparison views with the underlying variation policy. They should engineer tightly related sets where non-target dimensions are frozen, delay entry into comparison views until there is enough spread along at least one dimension, and overlay explicit "what changed" markers (for example, *Arm: track → pillow*). Ordering axes by dispersion can further surface informative contrasts while hiding unhelpful noise.

6.3 Rethinking low fidelity in a world of cheap high fidelity

Participants were split on sketch-like outputs (Theme 3). Classic prototyping work argues that low-fidelity artifacts reduce commitment and invite broad exploration, whereas high fidelity supports specification and judgment [11, 85, 102]. Data from this study show both dynamics, but in a GenAI context where high-fidelity images are nearly as easy to generate as sketches.

For some participants (P6, P8, P14), sketches did what the literature predicts: they made flaws obvious, encouraged lateral moves, and felt safer to experiment with before "spending" model calls on detailed renders. For others (P4, P11, P15), sketches obscured important evaluative cues. These participants reported difficulty imagining the sofa in a real room until they saw polished images, and some found repeated sketch iterations disorienting or discouraging. When sketches also failed to reflect targeted edits (P1, P3), people questioned whether the system had respected their intentions at all.

Taken together, these mixed reactions suggest that in an era where high-fidelity generation is cheap, the blanket advice to "start with low-fi" needs a more nuanced reading. Low-fi still appears valuable for mitigating premature fixation and inviting critique, but these findings highlight how novices can struggle when asked to evaluate only from sketches. A staged approach that responds to both task intent and user preference may be more appropriate than a fixed fidelity policy.

Design implication. Fidelity policies should be coupled to both task intent and user control. Interfaces should make fidelity levels visible, adjustable, and reversible, and should preserve provenance across levels so users can see that only their chosen dimensions changed. This can preserve the creative flexibility associated with low-fi while still giving novices enough detail to evaluate options. Larger, controlled studies are needed to test how different fidelity policies affect exploration breadth, fixation, and judgment in GenAI workflows.

6.4 Confidence, ownership, and perceived agency

RQ2 asked how participants felt about using *VariationWeaver* for product design. Confidence increased when the workflow scaffolded next steps and made progress feel incremental (P3, P4), and when the simulated colleague packet provided labeled exemplars to react to (P6, P10, P12; Theme 4). Prior work on parallel prototyping suggests that seeing multiple alternatives can support integration, reduce evaluation anxiety, and build a sense of direction [21, 22]. Participants' qualitative accounts are consistent with this: several described feeling "less pressure" once they could compare their designs with another person's labeled set rather than facing a blank canvas alone.

Ownership followed a similar pattern (Theme 5). Participants reported stronger authorship when they saw their intent reflected in the images or could gradually shape results through small, visible moves. This aligns with work on external representations and agency, where incremental, inspectable changes help people maintain a sense of control over complex systems [1, 86]. Conversely, ownership eroded when the model ignored targeted edits or produced outputs that felt distant from participants' mental images. In those cases, people described the sofas as "AI-generated" rather than "their design," even if they had supplied the original brief.

Design implication. To support confidence and ownership, GenAI interfaces should emphasize transparent, reversible moves that visibly link user actions to system responses, and they should make room for socially grounded comparison rather than relying solely on solitary prompting.

6.5 Limitations

This lab study focuses on novices (N=15) and a single domain (sofas), which may limit generality. The *asynchronous-by-proxy* design approximates, but does not replicate, multi-party handoffs in real studios. The study does not include a comparative baseline; the ranking probe is descriptive rather than inferential. Model non-responsiveness sometimes broke the intended structure of tightly related variant sets, confounding how participants experienced comparison and control. Finally, the study does not directly measure fixation or coverage, so mechanistic claims about search patterns are interpretive.

One practical limitation concerns how the system checks whether the image model actually followed users' constraints. In this prototype, adherence is monitored qualitatively through tags, logs, and inspection rather than via a second automated checker. More robust automatic adherence checking—for example, using a secondary model to verify that specific dimensions changed as requested—could reduce ambiguity, but would introduce additional latency and cost. We view such mechanisms as an important direction for future refinement rather than part of the current deployment.

6.6 Future Work

Comparative, fidelity-aware evaluation. A natural next step is a larger-scale, quantitative study that tests the speculative mechanisms suggested by this study's qualitative data: that intentional variation over tagged dimensions, structured comparison views, and staged fidelity jointly affect exploration breadth, verification cost, convergence quality, and perceived authorship. Such work would benefit from comparative baselines, for example contrasting *VariationWeaver* with a palette-only interface in the spirit of DesignWeaver and a standard text-to-image prompt box, while also systematically varying fidelity policies (e.g., sketch-first, render-first, and user-controlled) [85, 102]. The present findings surface the tensions and design parameters; future experiments can more precisely estimate effect sizes and boundary conditions once high-fidelity generation is cheap and routine.

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 [34, 110].

7 CONCLUSION

VariationWeaver builds on palette-based language scaffolding by coordinating three mechanisms—intentional variation over tagged dimensions, structured comparison views, and staged fidelity— to help novices explore and converge with text-to-image models. In a qualitative study with 15 novice designers, we observed three main patterns: (1) intentional variation over tagged dimensions helped participants notice, name, and revisit key design dimensions while keeping changes incremental and tractable; (2) structured comparison supported deliberate trade-offs when variants differed in clear, meaningful ways but felt redundant or confusing when images were too similar or models ignored requested changes; and (3) sketch-first outputs sometimes sparked creative exploration yet often left novices wanting higher-fidelity images when making final judgments in a context where polished renders are easy to generate. Taken together, these findings suggest that language scaffolds, variation policies, comparison layouts, and fidelity strategies need to be co-designed: simple "low-fi first" heuristics are not sufficient once high-fidelity generation is routine. We see opportunities for comparative, fidelity-aware evaluations and richer multi-person asynchronous workflows that test how these coordinated mechanisms affect exploration breadth, verification cost, convergence quality, and authorship. More broadly, embedding intentional variation, structured comparison, and staged fidelity into creative AI tools can help turn fast images into teachable, decision-ready design representations rather than isolated prompt responses.

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] Jonathan Chen and Dongwook Yoon. 2024. Exploring the Diminishing Allure of Paper and Low-Fidelity Prototyping Among Designers in the Software Industry: Impacts of Hybrid Work, Digital Tools, and Corporate Culture. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–14.

[15] 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

[16] 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

[17] 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.

[18] 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

[19] 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.

[20] 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

[21] 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

[22] 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

[23] K. J. Kevin Feng, Tony W Li, and Amy X. Zhang. 2023. Understanding Collaborative Practices and Tools of Professional UX Practitioners in Software Organizations. 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 764, 20 pages. doi:10.1145/3544548.3581273

[24] 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

[25] 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.

[26] William W. Gaver, Jacob Beaver, and Steve Benford. 2003. Ambiguity as a resource for design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 233–240. doi:10.1145/642611.642653

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

[28] 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

[29] Elena L. Glassman, Jeremy Scott, Rishabh Singh, Philip J. Guo, and Robert C. Miller. 2015. OverCode: Visualizing Variation in Student Solutions to Programming Problems at Scale. *ACM Trans. Comput.-Hum. Interact.* 22, 2, Article 7 (March 2015), 35 pages. doi:10.1145/2699751

[30] Asanshay Gupta, Vishnu Sarukkai, and Kayvon Fatahalian. 2025. Axes-and-Tags: LLM-Driven Design Galleries for Generative Content. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1975–1984.

[31] 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.

[32] Evans Xu Han, Alice Qian Zhang, Haiyi Zhu, Hong Shen, Paul Pu Liang, and Jane Hsieh. 2025. POET: Supporting Prompting Creativity and Personalization with Automated Expansion of Text-to-Image Generation. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 162, 18 pages. doi:10.1145/3746059.3747710

[33] Björn Hartmann, Loren Yu, Abel Allison, Yeonsoo Yang, and Scott R. Klemmer. 2008. Design as exploration: creating interface alternatives through parallel authoring and runtime tuning. In *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology* (Monterey, CA, USA) *(UIST '08)*. Association for Computing Machinery, New York, NY, USA, 91–100. doi:10.1145/1449715.1449732

[34] 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

[35] 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.

[36] Jonathan HG Hey, Caneel K Joyce, and Sara L Beckman. 2007. Framing innovation: negotiating shared frames during early design phases. *Journal of Design Research* 6, 1-2 (2007), 79–99.

[37] 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).

[38] 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.

[39] 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

[40] 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

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

[42] 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

[43] Gabe Johnson, Mark D Gross, Jason Hong, Ellen Yi-Luen Do, et al. 2009. Computational support for sketching in design: a review. *Foundations and Trends® in Human–Computer Interaction* 2, 1 (2009), 1–93.

[44] Tae Soo Kim, DaEun Choi, Yoonseo Choi, and Juho Kim. 2022. Stylette: Styling the Web with Natural Language. 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 5, 17 pages. doi:10.1145/3491102.3501931

[45] Tae Soo Kim, Yoonjoo Lee, Minsuk Chang, and Juho Kim. 2023. Cells, Generators, and Lenses: Design Framework for Object-Oriented Interaction 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 4, 18 pages. doi:10.1145/3586183.3606833

[46] 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.

[47] Maaike Kleinsmann and Rianne Valkenburg. 2008. Barriers and enablers for creating shared understanding in co-design projects. *Design studies* 29, 4 (2008), 369–386.

[48] 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

[49] 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).

[50] 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

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

[52] 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

[53] 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.

[54] 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

[55] 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.

[56] 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

[57] 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

[58] Jiaju Ma, Chau Vu, Asya Lyubavina, Catherine Liu, and Jingyi Li. 2025. Computational Scaffolding of Composition, Value, and Color for Disciplined Drawing. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 161, 15 pages. doi:10.1145/3746059.3747605

[59] Wendy E. Mackay and Michel Beaudouin-Lafon. 2025. Interaction Substrates: Combining Power and Simplicity in Interactive Systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 687, 16 pages. doi:10.1145/3706598.3714006

[60] 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.

[61] 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

[62] Nicolai Marquardt, Asta Roseway, Hugo Romat, Payod Panda, Michel Pahud, Gonzalo Ramos, Steven M. Drucker, Andrew D. Wilson, Ken Hinckley, and Nathalie Riche. 2025. ImaginationVellum: Generative-AI Ideation Canvas with Spatial Prompts, Generative Strokes, and Ideation History. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 159, 19 pages. doi:10.1145/3746059.3747631

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

[64] Justin Matejka, Michael Glueck, Erin Bradner, Ali Hashemi, Tovi Grossman, and George Fitzmaurice. 2018. Dream Lens: Exploration and Visualization of Large-Scale Generative Design Datasets. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3173943

[65] Michael McCurdy, Christopher Connors, Guy Pyrzak, Bob Kanefsky, and Alonso Vera. 2006. Breaking the fidelity barrier: an examination of our current characterization of prototypes and an example of a mixed-fidelity success. 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, 1233–1242. doi:10.1145/1124772.1124959

[66] 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.

[67] 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.

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

[69] 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.

[70] 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

[71] 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

[72] 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

[73] 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).

[74] Gaurav Parmar, Or Patashnik, Daniil Ostashev, Kuan-Chieh Wang, Kfir Aberman, Srinivasa Narasimhan, and Jun-Yan Zhu. 2025. Scaling Group Inference for Diverse and High-Quality Generation. *arXiv preprint arXiv:2508.15773* (2025).

[75] 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.

[76] 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.

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

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

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

[80] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.

[81] 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

[82] 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

[83] 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.

[84] 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

[85] 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.

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

[87] Donald A Schön. 1992. Designing as reflective conversation with the materials of a design situation. *Knowledge-based systems* 5, 1 (1992), 3–14.

[88] 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

[89] Nikhil Sharma and George Furnas. 2009. Artifact usefulness and usage in sensemaking handoffs. *Proceedings of the American Society for Information Science and Technology* 46, 1 (2009), 1–19.

[90] 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

[91] 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).

[92] 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

[93] 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

[94] 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

[95] 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

[96] Masaki Suwa and Barbara Tversky. 1997. What do architects and students perceive in their design sketches? A protocol analysis. *Design studies* 18, 4 (1997), 385–403.

[97] 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

[98] 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

[99] Maryam Tohidi, William Buxton, Ronald Baecker, and Abigail Sellen. 2006. User sketches: a quick, inexpensive, and effective way to elicit more reflective user feedback. In *Proceedings of the 4th Nordic Conference on Human-Computer Interaction: Changing Roles* (Oslo, Norway) *(NordiCHI '06)*. Association for Computing Machinery, New York, NY, USA, 105–114. doi:10.1145/1182475.1182487

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

[101] 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

[102] 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.

[103] Sitong Wang, Samia Menon, Dingzeyu Li, Xiaojuan Ma, Richard Zemel, and Lydia B Chilton. 2025. Schemex: Interactive Structural Abstraction from Examples with Contrastive Refinement. *arXiv preprint arXiv:2504.11795* (2025).

[104] Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J Cai. 2022. PromptChainer: Chaining Large Language Model Prompts through Visual Programming. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 359, 10 pages. doi:10.1145/3491101.3519729

[105] 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

[106] 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

[107] Yu-Chun Grace Yen, Jane L. E, Hyoungwook Jin, Mingyi Li, Grace Lin, Isabelle Yan Pan, and Steven P. Dow. 2024. ProcessGallery: Contrasting Early and Late Iterations for Design Principle Learning. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 112 (April 2024), 35 pages. doi:10.1145/3637389

[108] 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.

[109] 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

[110] 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 $\{\texttt{Dimension: Tag}\}^{\times \texttt{targetCount}}$.

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

- **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).

A.12 Stage Fidelity (Style Ramp) Prompts

- **Goal:** Automatically adjust image generation style (fidelity stage) based on number of user-selected tags to provide progressive refinement from concept sketches to final product photography.
- **Model(s):** `gpt-image-1` (all stages); `flux-schnell` (fallback for 0–4 tags).
- **Style Ramp Logic:** The system selects one of three fidelity stages based on `userSelectedTagCount` (count of user-selected tags, excluding AI-generated tags):
  - **Stage 1 (0–2 tags):** Quick Sketch (Black & White) — low fidelity, concept exploration.
  - **Stage 2 (3–4 tags):** Colored Sketch — medium fidelity, design refinement.
  - **Stage 3 (5+ tags):** Realistic (Professional Photography) — high fidelity, final product.
- **Messages:** Full prompt construction depends on stage and presence of user text:

  - **Stage 1 (0–2 tags):** If user prompt empty: "a single sofa with `tagDescription`. Quick sketch style, black and white line drawing, minimalist, rough pencil sketch, concept art, simple lines, no shading, no color, white background, sketchy and loose, artistic rough draft, design concept". If user prompt present: "`userPrompt` with `tagDescription`. [same style suffix]".
  - **Stage 2 (3–4 tags):** If user prompt empty: "a single sofa with `tagDescription`. Colored sketch style, hand-drawn illustration, artistic sketch with color, watercolor-like, loose brush strokes, artistic rendering, concept art with color, sketchy but colorful, design illustration, white background". If user prompt present: "`userPrompt` with `tagDescription`. [same style suffix]".
  - **Stage 3 (5+ tags):** If user prompt empty: "a single sofa with `tagDescription`. Show only one sofa, no multiple sofas. Pure white background, absolutely no color artifacts, no shadows, no gradients, no bubbles, no smudges, no background elements. Studio lighting with clean, crisp edges. Professional product photography on pure white backdrop". If user prompt present: "`userPrompt` with `tagDescription`. [same style suffix]".
- **Inputs:** `userSelectedTagCount` (integer); `tagDescription` (NL phrase from all tags, user-selected + AI-generated); optional `userPrompt` (user text).
- **Outputs:** Full prompt string with stage-appropriate style suffix; `imageStyle` metadata field ("Quick Sketch (Black & White)", "Colored Sketch", or "Realistic (Professional Photography)").
- **Note:** Stage selection is based solely on `userSelectedTagCount`, not total tag count (which includes AI-generated tags). This ensures user intent drives fidelity level.

A.13 Automatic Tag Generation Logic

- **Goal:** Automatically generate complementary dimension–tag pairs to enrich sparse user selections, improving image quality and design exploration.
- **Model(s):** `gpt-5-nano` (via `generateAttributesForDimensions`).
- **Generation Rules:** Based on `userSelectedTagCount`, the system generates additional tags before image generation:
  - **0 tags selected:** Generate 3 new dimension–tag pairs (fully AI-driven exploration).
  - **1–2 tags selected:** Generate 2 additional dimension–tag pairs (complementary attributes).
  - **3–7 tags selected:** Generate 1 additional dimension–tag pair (fine-tuning).
  - **8+ tags selected:** No additional tags (user has sufficient specification).
- **Messages: user** — "Generate `targetCount` diverse dimension–tag pairs for a sofa design. Existing tags: `existingTags`. User prompt: `userPrompt`. Return JSON object {Dimension: Tag}."
- **Inputs:** `userSelectedTagCount`; `existingTags` (user-selected tags); optional `userPrompt`; `targetCount` (1, 2, or 3 based on rules above).
- **Outputs:** `additionalTagsGenerated` (JSON object of new dimension–tag pairs); `finalTags` (merged user-selected + AI-generated tags, used for `tagDescription` in image generation).
- **Note:** Generated tags are stored separately in `inferred_dimension_tags` for tracking; only user-selected tags count toward stage fidelity selection (Section A.12).

A.14 High-fidelity Style Consistency for Sweeps & Compare

- **Goal:** Ensure visually comparable candidates in structured exploration views by standardizing fidelity for sweep grids (Section A.3) and compare-mode generation (Section A.11).
- **Model(s):** `gpt-image-1` (primary); `flux-schnell` (previews/fallback), consistent with Sections A.2 and A.3.
- **Style Policy (Stage 3):** Both sweeps and compare-mode use the same high-fidelity style suffix, regardless of `userSelectedTagCount`: "Show only one sofa, no multiple sofas. Pure white background, absolutely no color artifacts, no shadows, no gradients, no bubbles, no smudges, no background elements. Studio lighting with clean, crisp edges. Professional product photography on pure white backdrop."
- **Sweeps (Section A.3):** All cells in the $m \times n$ grid inherit the Stage 3 style suffix so differences are attributable to target dimensions, not inconsistent lighting or background.
- **Compare-mode (Section A.11):** Child candidates synthesized from multiple parents (or enforced `dimension:value`) also always use the Stage 3 style suffix, so visual comparison focuses on geometry and attributes rather than mixed fidelities.
- **Inputs/Outputs:** No new fields beyond those in Sections A.2, A.3, and A.11; this section specifies a global style policy applied at render time.

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. 12. 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. 13. 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 our design goals in §3.1. 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 tagged factors), *Comparison* (aligned, ranked side-by-side views), *Terminology* (naming and stabilizing factors via palettes), and *Fidelity* (staging 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	[57, 109]
N2 Structured Comparability	Unaligned alternatives; tab overload; trade-offs not explicit	Comparison	[13, 76, 77]
N3 Between-session Reflection	Momentum and rationale dissipate between turns	Terminology, Variation, Comparison	[40, 72, 106]
N4 Externalized Rationale	Intent and criteria remain implicit; poor handoff quality	Terminology	[48, 101, 110]
Post-GenAI (N5–N8)			
N5 Mesoscale Sensemaking*	Candidate overload; subtle differences missed in large batches	Variation, Comparison, Fidelity	[7, 28, 61, 94]
N6 Convergence with Diversity	Premature homogenization; loss of exploratory coverage	Variation, Comparison, Fidelity	[21, 22, 25]
N7 Teachable Dimensions	Tacit criteria unnamed; terms drift across collaborators	Terminology	[20, 52, 88, 94, 106]
N8 Shareable, Replayable Strategy	Exploration strategy cannot be transferred or reproduced later	Terminology, Variation, Comparison	[18, 19, 82, 105]

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

Table 3. How needs (N1–N8) map to design goals (DG1–DG3). DG1: intentional variation over tagged dimensions; DG2: structured 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 (intentional variation over tagged dimensions) gives novices controllable levers for changing designs while preserving context, which helps surface tacit criteria (N4, N7), improves awareness and transfer of strategy across turns (N1, N8), and sustains between-session reflection (N3). DG2 (structured comparison) makes alternatives commensurate and trade-offs explicit (N2), supports mesoscale sensemaking when many candidates accumulate (N5),

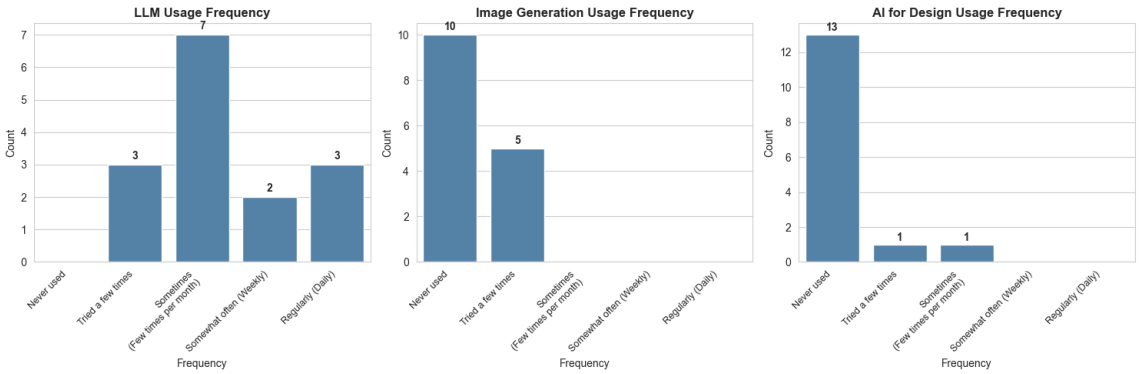

Fig. 14. Self-reported prior exposure to AI tools before the study (N=15). Panels show frequency categories for three tool types: large-language-model (LLM) use, text-to-image use, and "AI for design." Values are counts.

and enables principled convergence without collapsing diversity (N6), while also helping maintain momentum across turns (N3, N8). DG3 (staged fidelity) aids perception and evaluation at scale (N5) and complements convergence (N6) by preserving breadth early and only sharpening visual detail as decisions stabilize.

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; revised December 04, 2025