

FocusGEN: Expanding Visual Design Exploration with a Simulated Focus Group of Persona Agents

Jaewon Choi
Hanwha Life
Seoul, South Korea
jaewonch@hanwha.com

Helena Vasconcelos
Stanford University
Palo Alto, California, USA
helenav@stanford.edu

Hyun Lee
KAIST
Daejeon, South Korea
hyunini@kaist.ac.kr

Carolyn Zou
Stanford University
Palo Alto, California, USA
cqz@stanford.edu

Tak Yeon Lee
KAIST
Daejeon, South Korea
takyeonlee@kaist.ac.kr

Michael Bernstein
Stanford University
Palo Alto, California, USA
msb@stanford.edu

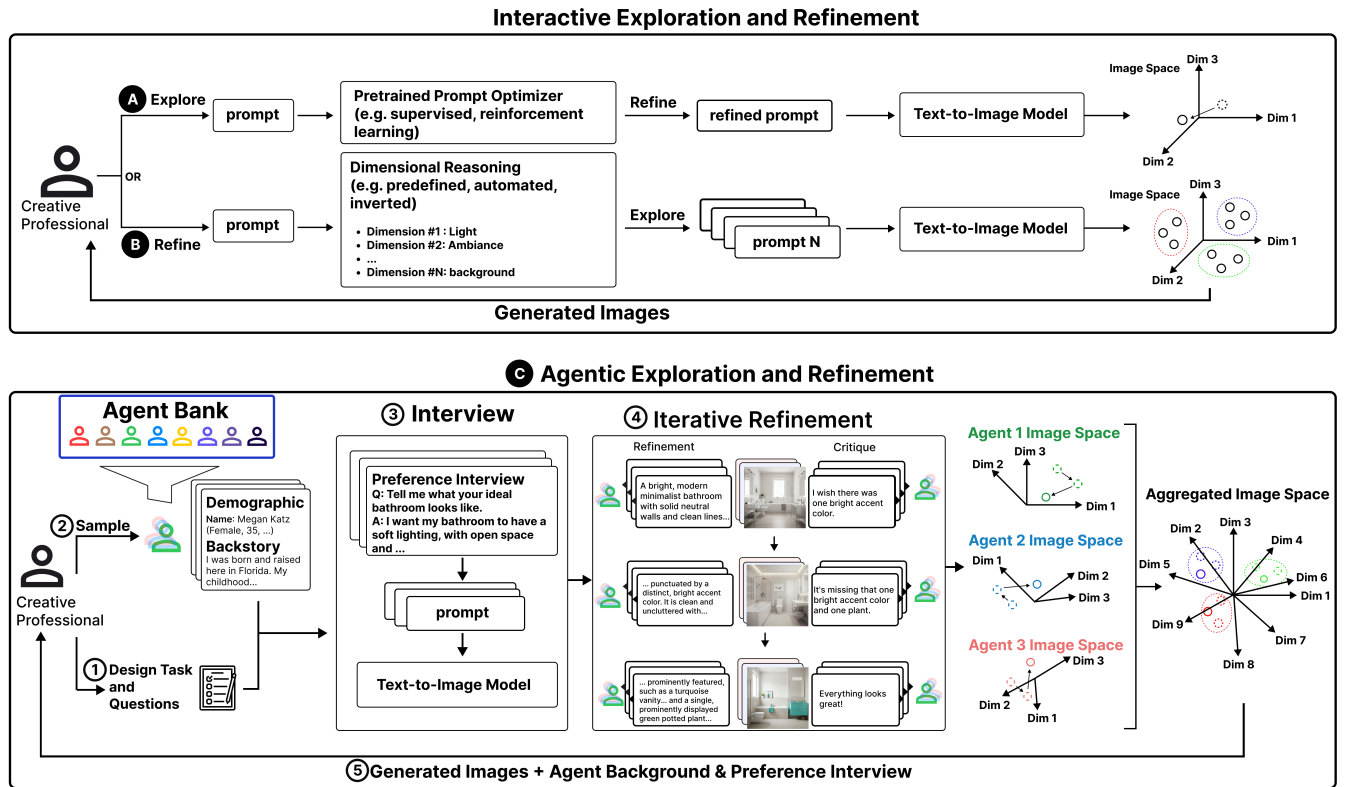


Figure 1: Comparison of Generative Exploration Paradigms. (Top) Current Approaches: (A) Prompt Optimization requires training, which is computationally intensive and difficult to scale. (B) Dimensional Reasoning relies on user-defined axes (e.g., Lighting, Ambiance), confining exploration to what the designer already knows to look for. (Bottom) Our Approach: (C) FocusGEN introduces external perspectives through persona agents equipped with demographics, backstories, and aesthetic preferences. By conducting open-ended preference interviews, the system elicits dimensions the designer may not have considered. Each agent independently generates visual concepts conditioned on its distinct profile, surfacing diverse directions—the designer’s blind spots—to support creative professionals in divergent exploration.

Abstract

Creative professionals rarely design for themselves—they design for audiences whose preferences they must anticipate. Yet current text-to-image exploration tools derive diversity entirely from the designer’s own input—their prompts, their chosen dimensions, their

search queries—confining exploration to what the designer already knows to look for. We present FocusGen, an interactive system that introduces external perspectives into visual design exploration through a “virtual focus group” of simulated persona agents. In contrast to prior persona systems in which multiple agents converge as critics on a single evolving artifact, FocusGen uses personas

as parallel generators: each agent—constructed from demographic data, a procedurally generated backstory, and aesthetic preferences elicited through interviews—independently drives an iterative generation loop that produces its own visual concept, transforming one design brief into a spectrum of audience-conditioned directions. We evaluate the approach in three parts. With real human participants, we confirm that the iterative refinement loop produces outputs people prefer over zero-shot generation. With synthetic agents at scale, we show that persona conditioning yields higher visual diversity than a generic-assistant baseline—measured by CLIP distance and corroborated by human perceptual judgments—and that open-ended preference interviews yield more diverse outputs than structured ones for both human and synthetic cohorts, while also revealing that agent cohorts recover only part of the diversity of comparable human cohorts. A qualitative study with 16 creative professionals suggests FocusGen helps designers discover unanticipated directions, overcome fixation, and probe audience contexts—while surfacing stereotyping risks that we analyze. We position FocusGen as a divergence scaffold for early-stage ideation rather than a substitute for audience research.

CCS Concepts

• **Human-centered computing** → User studies; Empirical studies in interaction design; • **Computing methodologies** → Computer vision.

Keywords

Simulated Focus Groups, Persona Agents, Design Exploration, Creativity Support Tools, Text-to-Image Generation

1 Introduction

Creative professionals rarely design in a vacuum. Whether crafting a brand identity, a product concept, or a marketing campaign, their work must resonate with audiences whose preferences differ from their own. In the early stages of design, exploring diverse visual directions is critical—yet this is precisely where current text-to-image tools fall short. Not because the models lack capability, but because the exploration process is fundamentally bounded by the designer’s own perspective.

Today’s exploration tools all share a common limitation: they derive diversity from the *user’s* input. Prompt engineering [33, 52] relies on the designer to articulate the right modifiers. Dimensional reasoning tools like Luminare [45] and DesignAID [8] help users navigate semantic axes, but those axes are extracted from the user’s own prompts or generated images. Even automated prompt optimization [20, 50] refines what the user has already expressed. These approaches can surface “known unknowns”—variations the designer suspects exist but hasn’t yet explored—but they cannot surface what the designer has not thought to look for: the *unknown unknowns*.

In professional practice, these unknown unknowns are traditionally surfaced through audience feedback—focus groups, user interviews, and design critiques that bring external perspectives into the process. But gathering such feedback is expensive, slow, and impractical during early-stage ideation when concepts are still

forming. We ask: can simulated audience perspectives serve this role?

We present **FocusGen**, an interactive system that introduces external perspectives into visual design exploration through a “virtual focus group” of persona agents. Prior work has shown that LLM-based agents grounded in demographics and rich backstories can produce idiosyncratic, individuated responses rather than generic stereotypes [32, 34, 36]. FocusGen builds on this foundation by adding a layer of **aesthetic preference elicitation**: each agent is interviewed to surface its specific visual priorities, grounded in its life context. Each agent’s elicited preferences are supplied to an iterative generation and critique loop. This workflow is intended to make alternatives interpretable through stated persona preferences. Our evaluations test whether heterogeneous persona-profile conditioning increases output diversity, but they do not establish that outputs are discriminably faithful to individual personas or representative of demographic groups.

FocusGen constructs its agents synthetically: demographic profiles are sampled from census data [9], enriched with procedurally generated backstories [32], and grounded through open-ended interviews that the agents answer in character. When given a design brief, each agent independently generates visual concepts conditioned on its distinct profile. The system parallelizes this across the full agent cohort, producing a gallery of audience-specific visual concepts. To validate this approach, we first confirmed with real human participants that the refinement loop produces outputs people prefer over zero-shot generation, then showed with synthetic agents at scale that persona diversity translates into measurable visual diversity.

In summary, the contributions of this work are:

- **FocusGen**: A parallelized-divergent interaction model for persona-based creativity support—persona agents as independent, parallel generators of audience-conditioned design alternatives, rather than convergent critics of a single evolving artifact—instantiated in FocusGen with a 1,000-agent, demographically grounded Agent Bank and an interface that shifts the designer’s role from prompt engineering to audience curation.
- **Empirical evidence** that (i) the iterative refinement loop produces outputs that human participants prefer over zero-shot generation, (ii) persona conditioning measurably and perceptibly increases the visual diversity of generated alternatives over a generic-assistant baseline, and (iii) open-ended preference interviews expand output diversity relative to structured interviews for both human and synthetic cohorts—together with an explicit characterization of what these evaluations do not establish (discriminative persona faithfulness and demographic representativeness) and of a persistent human-agent diversity gap.
- **The design insight** that preference interviews—where designers craft questions for simulated personas rather than prompts for a model—serve as a productive interaction paradigm for divergent exploration, shifting the designer’s role from commanding to curating, with qualitative evidence from 16 creative professionals on how this paradigm functions in practice.

2 Related Work

2.1 Designing for Audiences: The Need for External Perspectives

Professional designers rarely design for themselves—they design for audiences whose preferences, contexts, and reactions they must anticipate. This is well-established in the design literature: designers are inherently not their users, and deliberate methods are needed to bridge this empathy gap [5, 26]. In practice, external feedback—from focus groups, user interviews, and design critiques—provides genuinely different information than what designers generate internally [30]. Systems like Voyant [53] demonstrated that even non-expert crowd workers, when selected to match a target audience demographic, can provide perception-oriented feedback that meaningfully changes design outcomes. Furthermore, creating and comparing multiple design alternatives in parallel leads to better results and greater divergence than sequential refinement [16]—a principle FOCUSGEN enacts by parallelizing feedback across a diverse agent population.

However, gathering real audience feedback is expensive, slow, and impractical during early-stage ideation when concepts are still forming. This motivates the core question of our work: can *simulated* audience perspectives serve this role?

2.2 Current Exploration Tools Confine Discovery to Known Unknowns

2.2.1 Design Fixation and AI-Induced Homogenization. Design fixation—the unconscious adherence to a limited set of ideas that restricts exploration [12, 24]—is a well-documented challenge that affects novices and experts alike [13]. Recent work has shown that generative AI tools can *exacerbate* rather than mitigate this problem. Wadinambiarachchi et al. [49] found that AI image generators increase fixation compared to traditional inspiration sources, producing fewer ideas with less variety. At the collective level, Doshi and Hauser [15] demonstrated that while LLM access improves individual creative output, it significantly reduces the *diversity* of ideas across users—a homogenization effect confirmed across multiple LLMs [23]. This creates a paradox: tools designed to support creativity may narrow the very design space they aim to expand.

2.2.2 Prompt-Based Exploration and Its Limits. The primary means of controlling text-to-image models remains the text prompt, and prompt engineering is difficult [33, 52]. Users explore prompts opportunistically rather than systematically, constrained by what they can articulate and imagine [54]. Subramonyam et al. [44] formalized this as the “Gulf of Envisioning”—users can only prompt for what they can envision, creating a cognitive bottleneck that bounds exploration to *known unknowns*.

Automated prompt optimization [20, 50] and iterative refinement frameworks like Self-Refine [31] and reflexion [41] reduce this burden by using LLMs to critique and revise outputs. However, these approaches still optimize for the user’s stated intent using a generic critic—not the preferences of a specific audience.

2.2.3 Dimensional Reasoning Tools. Recent creativity support tools facilitate exploration through *dimensional reasoning*—navigating

the design space via semantic axes. Luminate [45] generates dimensions for users to adjust, DesignAID [8] generates diverse ideas from user text, DesignWeaver [46] extracts dimensions from design documents, and POET [19] discovers dimensions of homogeneity via prompt inversion. While valuable, these tools share a fundamental limitation: their dimensions are derived from the user’s own input, confining exploration within the user’s existing conceptual frame. They help designers explore “known unknowns” but cannot surface the *unknown unknowns*—perspectives the designer has not thought to look for [25].

FOCUSGEN addresses this gap by introducing *external* perspectives into the generation loop. Unlike prompt recommendation systems [7, 17, 43] that optimize for the user’s predicted intent, we simulate a diverse audience whose feedback may contradict or expand the designer’s initial framing, supporting divergent ideation rather than convergence.

2.3 Simulating Populations with Persona Agents

The development of generative agents that simulate specific people offers a path toward scalable audience-like feedback. Park et al. [34] demonstrated that LLMs can generate believable, individualized behaviors, and Social Simulacra [35] showed that designers used simulated community members to see “beyond the social interactions that they intend their design to produce.” These agents can be grounded in demographic data [3, 9], rich narrative contexts [32], or specific user preferences [10, 37], maintaining character consistency across tasks [39, 47].

However, the reliance on demographic conditioning necessitates caution regarding *representational harms* [28, 51]. Park et al. [36] demonstrate that enriching agents with specific life histories substantially mitigates stereotyping by shifting the model from broad generalizations to individual traits.

Several recent systems apply persona agents to creative and intellectual work, and they do more than post-hoc critique. Poster-Mate [40] lets audience persona agents propose edits to text, imagery, and theme that are integrated directly into the poster canvas; Proxona [10] supports creators in making sense of their audience and refining early-stage creative work through conversations with data-grounded audience personas; PersonaFlow [29] uses simulated domain experts to drive literature retrieval, critique, and the generation of revised research questions; and SimTube [22] generates simulated audience commentary on video. We therefore do not position FOCUSGEN as the first system to bring personas into a generative loop. The distinction we draw is one of interaction topology. These systems are predominantly *convergent*: multiple personas critique, annotate, or refine a single evolving artifact that the creator owns. FOCUSGEN is *parallel-divergent*: each persona independently authors its own visual concept end-to-end—preference elicitation, prompt construction, and iterative refinement—and the system’s unit of output is the resulting spectrum of audience-conditioned artifacts rather than an improved single artifact. This topological choice is motivated by evidence that developing alternatives in parallel produces more divergent outcomes than sequentially refining one concept [16]; parallelizing across personas is what

allows population-level diversity to translate into visual diversity, the property we test in Evaluations 1 and 2.

A complementary line of work uses multi-agent coordination to improve generative image creation itself. CREA [48] orchestrates agents in creative-process roles—such as a creative director, critics, and editors—that collaborate on editing and refining images. The axis of agent differentiation differs from ours: CREA’s agents embody stages of the creative process and converge on one high-quality output, whereas FOCUSGEN’s agents embody members of a simulated audience and diverge into many outputs. The two approaches are composable—CREA-style role specialization could strengthen the per-agent refinement loop inside each FOCUSGEN persona—but they target different problems: making one image better versus exposing a designer to directions they did not think to look for.

FOCUSGEN’s individual components—demographic persona construction [9], narrative grounding [32], and iterative self-refinement [31, 41]—each build on established techniques. The contribution lies in their composition into the parallelized-divergent workflow above: using each persona’s elicited preferences to drive independent generation, parallelizing across a demographically diverse agent population, and providing an interface that lets designers act as curators of a simulated audience. Whether this translation of population diversity into visual diversity actually occurs is an empirical question, which we test in Section 4.

3 FOCUSGEN: System Design

This section describes FOCUSGEN, an interactive system that introduces synthetic external audience perspectives into text-to-image design exploration. We first describe the designer-facing workflow—how a designer configures, launches, and explores results—then detail the two underlying mechanisms that power it: persona initialization and iterative image generation.

Implementation Details. For all experiments and system instantiations described in this paper, FOCUSGEN was implemented using Google’s *gemini-2.5-flash* Large Language Model for agent simulation and critique, and Google’s *imagen-3.0-generate* text-to-image model for visual generation.

3.1 Designer Workflow

FOCUSGEN provides designers with access to a diverse “Agent Bank”—a repository of 1000 persona agents, each initialized with demographic data [9] and a procedurally generated backstory [32]. Because each agent brings a distinct perspective grounded in its unique life context, the system can transform a single design brief into a gallery of audience-specific visual directions. The designer curates this process through a two-stage workflow.

3.1.1 Stage 1: Task Creation. The designer first selects a cohort of agents using demographic filters (age, gender, education, income) to target specific audience segments. They then define the creative brief: a high-level **Design Objective** and a set of **preference interview questions** (see Table 1). The interview step is critical—it requires designers to shift from *commanding* an AI to *interviewing* a persona. Questions can range from specific and orthogonal (e.g., separating “lighting” from “materials”) to broad and open-ended

(e.g., “Describe your five most important preferences”). As we show in Evaluation 2, this choice significantly affects the diversity of the resulting outputs. Once configured, the system generates visual concepts for each agent in parallel.

3.1.2 Stage 2: Feedback Exploration. The **Feedback Exploration View** (Figure 2) displays a gallery of the final images from all agents, each tagged with the generating agent’s demographics. Clicking any image reveals a **detail view** showing the full iterative refinement history (with preference scores per iteration), the agent’s structured critiques, interview responses, and complete profile. This allows designers to understand not only *what* was generated but *why*, connecting visual output to persona reasoning. Insights from one round inform subsequent experiments with refined filters, goals, or interview questions.

3.2 Persona Initialization

Behind the scenes, each agent in the Agent Bank is constructed through a two-layer process that establishes both a stable identity and task-specific aesthetic preferences.

3.2.1 Layer 1: Core Identity. The core identity combines structured demographic data (e.g., age, ethnicity, occupation) with a rich personal backstory (a first-person narrative detailing life events, hobbies, and memories). Demographic profiles are sampled from real-world datasets like the U.S. Census [9], and the backing LLM procedurally generates a unique, coherent backstory for each profile [32]. This procedural approach enables FOCUSGEN to construct the Agent Bank at scale without requiring real participant data. To validate the profile-to-image stage of this pipeline, we first tested it with preference profiles collected from real human participants (Technical Validation, Section 4.1) before applying it to synthetic agents (Evaluations 1 and 2).

3.2.2 Layer 2: Preference Elicitation. When the designer launches a task, each selected agent is prompted to answer the interview questions in character, inferring preferences from its demographic profile and backstory (e.g., an agent whose backstory centers on carpentry might cite “natural timber finishes”); we treat such responses as the agent’s operative preferences for generation, without claiming they mirror the preferences of real members of the corresponding group (Section 6.4). The output is a standardized textual *Persona Profile* that feeds into the image generation pipeline. Concretely, the agent prompt concatenates four components: the demographic profile (≈ 150 tokens), a procedurally generated backstory (≈ 1000 tokens), the elicited preference interview responses (≈ 400 tokens), and a roleplay instruction (≈ 200 tokens). A full example is provided in the Appendix (Table 4).

3.3 Iterative Image Generation

The final step translates each agent’s *Persona Profile* into a visual concept aligned with that profile. Because text-to-image models often miss details in complex prompts [1], a single-shot generation may not fully capture a persona’s preferences. The system therefore employs an iterative refinement loop [31, 41].

At each iteration, the agent critiques the generated image against its *Persona Profile*, identifying elements to preserve and deviations to correct, then generates a revised prompt. The agent maintains a

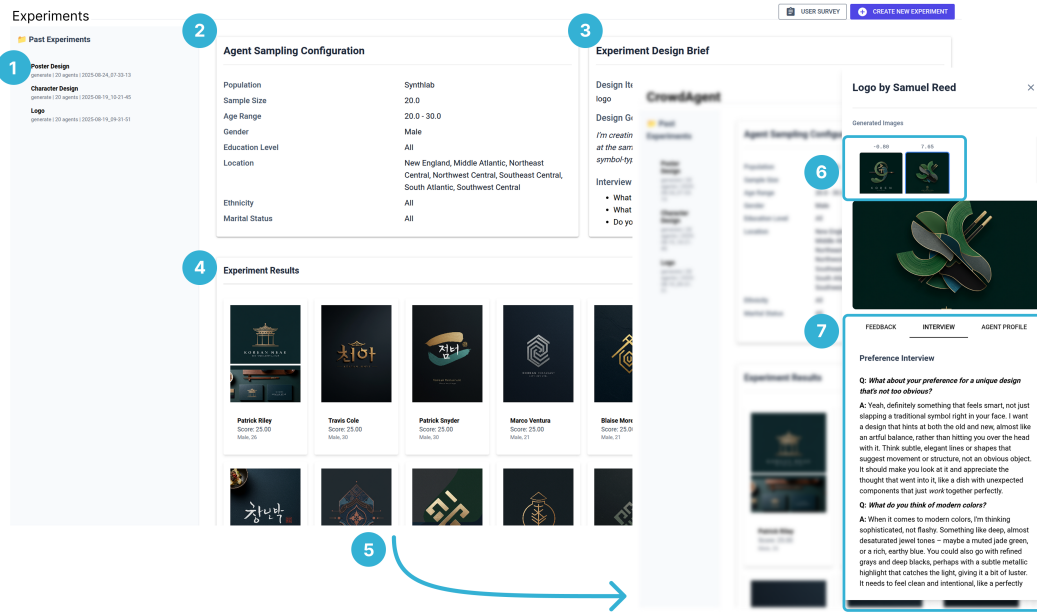


Figure 2: The Feedback Exploration View. (1) The navigation panel lists all past experiments for easy access. When an experiment is selected, its (2) agent configuration and (3) design brief are displayed. (4) The main gallery populates with a mood board of images from the persona crowd. (5) Clicking an image opens a detail view, revealing (6) the iterative refinement history, (7) the agent’s full textual feedback and interview responses, and its demographic profile.

Goal: Design the ideal bedroom.

Format A: Structured Questions

- 1 Ideal colors and materials for walls, floor, ceiling.
- 2 Ideal bed (shape, frame, bedding colors, textures).
- 3 Other furniture or items to include.
- 4 Type of lighting (ceiling, bedside, accent).
- 5 Window coverings (curtains, blinds, material).

Format B: Open-Ended Question

- 1 Describe five or more preferences for your bedroom in order of importance.

Table 1: Example interview questions for a bedroom design task. Designers craft their own questions, which can range from specific (Format A) to open-ended (Format B). This choice significantly affects output diversity (Evaluation 2).

history of past prompts and critiques to diagnose persistent failures. The loop terminates when no deviations remain or a maximum of $N = 5$ iterations is reached. Because the process is stochastic and non-monotonic, the agent reviews the entire image sequence and selects the best via pairwise TrueSkill ranking [21].

4 Evaluations

Our evaluations are designed to support three claims, and we are explicit about a fourth that they do not support. They show (1) that,

given a fixed preference profile grounded in a real person’s data, the iterative refinement loop produces outputs that the person prefers over zero-shot generation (Technical Validation, Section 4.1); (2) that conditioning generation on diverse personas yields measurably and perceptibly more diverse output sets than a generic-assistant baseline (Evaluation 1, Section 4.2); and (3) that open-ended preference elicitation yields more diverse outputs than structured elicitation, a pattern that holds for both human participants and synthetic agents (Evaluation 2, Section 4.3). They do not show (4) that a synthetic agent’s outputs are *discriminably faithful* to its specific persona—for example, that a participant would prefer images generated from their own profile over images generated from someone else’s—nor that synthetic agents’ self-generated preferences are representative of the demographic groups they are sampled from. Claim (4) requires a discriminative test that was not part of this work; we describe it in Section 6.4 and treat it as the most important next validation step. The qualitative study (Section 5) examines how creative professionals use the system in practice, independent of claim (4).

All studies were approved by the Institutional Review Board (IRB) at the institution where the study occurred.

4.1 Technical Validation: Iterative Refinement Produces More Preferred Outputs

Before testing whether diverse personas produce diverse outputs, we validate the generation pipeline itself: given a preference profile elicited from a real person, does the iterative loop produce

images that the person actually prefers? We emphasize the scope of this validation. It tests the profile-to-image stage of the pipeline, holding the preference profile fixed; it does not test the persona-to-preference stage (whether synthetic agents generate preferences representative of the people or groups they are modeled on), and, because it lacks a generic-critic control and a cross-persona comparison, it cannot by itself distinguish persona-specific alignment from generic image improvement. We recruited 27 participants from Prolific, collected their demographic profiles, backstories, and aesthetic preferences, and used this data to construct persona agents that served as digital proxies for each participant.

4.1.1 Method. Each session proceeded in three stages. First, the participant’s data was used to initialize a persona agent. Second, the system generated a sequence of up to five candidate images using the iterative refinement pipeline, terminating early if the agent identified no deviations. Third, both the participant and their corresponding agent independently ranked the full image sequence through pairwise comparisons (without knowledge of generation order), yielding TrueSkill scores [21] that track perceived quality across iterations.

4.1.2 Results.

Users Actively Prefer Refined Images. Participants selected a refined image (i.e., any image beyond the initial zero-feedback generation) as their top choice in the majority of sessions. To confirm this preference was not merely an artifact of there being more refined images (up to four) than initial images (one) in each session, we compared the observed choices to a random-selection baseline adjusted for this imbalance. A chi-squared goodness-of-fit test confirmed that users selected refined images significantly more often than this adjusted baseline ($\chi^2(1, N = 106) = 5.20, p < .05$), demonstrating that the preference is statistically robust and not simply a product of the number of options presented.

The Final Outcome is a Marked Improvement over the Initial Image. This preference is supported by a significant increase in alignment scores. A paired samples t-test confirmed that the average TrueSkill score of users’ final chosen image ($M = 15.57$) was more than double that of their initial, zero-feedback image ($M = 7.30, p < .001$). This large effect size demonstrates that the iterative refinement pipeline produces a meaningfully better final outcome.

Human-Agent Ranking Agreement. We also examined whether human and agent TrueSkill rankings of the full image sequence were correlated. The correlation was weak ($r = .26, p = .19$). Plausible contributors include differences in evaluation style—human evaluators judge holistically, weighing composition, mood, and coherence, while the agent critiques literal adherence to stated preferences—and the stochastic, non-monotonic nature of text-to-image refinement. We treat this result as a substantive caveat rather than an incidental one: a weak correlation means the agent’s judgment of “best for this profile” only partially tracks the profile owner’s own judgment. The preceding results therefore establish that refinement produces outcomes people prefer; they do not establish that an agent’s selections are discriminably aligned with its specific persona. The direct test—asking participants to choose between outputs generated from their own profile and outputs generated

from another participant’s profile, and including a generic-critic refinement arm—was not part of this study and is the most important next validation step (Section 6.4).

4.2 Evaluation 1: Agent Bank Diversity Translates to Visual Diversity

Having established that participants generally prefer refined outputs over the initial generation (Section 4.1), we now test whether the demographic and narrative diversity of our simulated agents translates into greater visual diversity. Using the Agent Bank that we constructed (U.S. census demographics, procedural backstories [9, 32]), we compared the collective output from a diverse agent cohort against outputs produced without personalization.

Conditions. For each task, we generated 3×3 grids of nine images under two conditions. In the **Agent-Enabled** condition, each image was generated by a unique Persona Agent from the Agent Bank. In the **Agent-Disabled** baseline, the persona layer was deactivated and a generic LLM assistant (prompt: “You are a helpful and creative assistant that can suggest new and inspiring designs. Given the task, think outside the box and provide imaginative and innovative suggestions”) served as the critic. Both conditions used the same iterative refinement process. We quantify diversity using average pairwise CLIP distance and CLIP dispersion. We note that CLIP distance measures semantic and visual dissimilarity but not design quality or relevance—random, off-brief images would also score high. We therefore complement it with a human perceptual evaluation.

Both conditions received the identical design brief and the identical refinement scaffold (the same critique-and-revise prompt template, maximum of $N = 5$ iterations, and TrueSkill selection step). The only difference was the identity supplied to the critic: in the Agent-Enabled condition, the critic received the full persona profile (demographics, backstory, elicited preferences, roleplay instruction; see Table 4); in the Agent-Disabled condition, it received only the generic-assistant prompt above.

Procedure. We recruited 20 participants from Prolific who viewed randomized pairs of grids side-by-side across five creative tasks and selected the more visually diverse one (or ‘Tie’; see Figure 7 in the Appendix).

4.2.1 Results. The **Agent-Enabled** condition yielded higher CLIP diversity across all five tasks (Table 3): **58%** average increase in pairwise distance and **33%** in dispersion, with the most dramatic gain in Interior Design (**180%**), likely because generic models produce especially homogeneous outputs in this domain (defaulting to modern minimalist styles). Participants also significantly preferred the Agent-Enabled grids as more visually diverse ($\chi^2(2, N = 468) = 86.81, p < .001$). Together, these results show that persona-profile conditioning produced grids with higher CLIP diversity that participants also perceived as more diverse than grids generated using a shared generic critic.

Limitations of This Comparison. The persona condition introduced substantially more and more heterogeneous conditioning text than the generic-critic condition, including demographic information, a procedurally generated backstory, elicited preference

Table 2: Open-ended interviews produce higher visual diversity for both humans and synthetic agents. Values show CLIP Distance with Dispersion in parentheses.

Task	Humans		Agents	
	Struct.	Open	Struct.	Open
UI Design	.27 (.10)	.34 (.14)	.24 (.10)	.29 (.12)
Packaging	.20 (.09)	.26 (.11)	.21 (.09)	.24 (.10)
Logo	.20 (.09)	.25 (.11)	.09 (.03)	.12 (.05)
Interior	.20 (.08)	.25 (.10)	.10 (.04)	.14 (.06)
Food	.13 (.05)	.21 (.10)	.09 (.03)	.14 (.06)
Avg	.20 (.08)	.26 (.10)	.15 (.06)	.19 (.08)

responses, and a roleplay instruction. This comparison therefore establishes a system-level diversity gain from the full persona-profile conditioning pipeline, but does not isolate which profile components caused that gain. In particular, the observed difference may reflect persona semantics, prompt length, lexical heterogeneity, preference variation, or a combination of these factors.

4.3 Evaluation 2: Open-Ended Interviews Expand Design Diversity

We now test the claim: that the way preferences are elicited affects the diversity of the resulting visual concepts.

4.3.1 Methodology. We compared Structured interviews against Open-Ended interviews across five creative tasks, evaluating both a **Human Cohort** (40 Prolific participants, $N_{open}=23$, $N_{struct}=17$) and an **Agent Cohort** (24 synthetic agents from the Agent Bank, interviewed using both formats). In the Structured condition, designers provide specific, orthogonal questions (e.g., separating “lighting” from “materials”). In the Open-Ended condition, agents receive a single broad prompt (e.g., “Describe your five most important preferences”), allowing them to self-define the dimensions that matter most. We quantify diversity using average pairwise CLIP distance and CLIP dispersion.

4.3.2 Results. Results from 115 user sessions and 120 agent sessions (Table 2) confirm that Open-Ended interviews produce higher diversity in both cohorts. For humans, Open-Ended increased CLIP Distance by **30%** and Dispersion by **25%**. For synthetic agents, the pattern was consistent (27% and 33% increases, respectively). Notably, the Open-Ended condition produced higher CLIP Distance in all five tasks for both cohorts. A sign test confirms this perfect consistency is unlikely by chance ($p = .031$, one-sided), though we note this test has minimal statistical power with only five data points—the result is driven entirely by the consistency of the pattern rather than the magnitude of individual effects. We therefore treat this as a *consistency check* rather than our primary evidence.

The Human-Agent Diversity Gap. Comparing the cohorts reveals a limitation we want to state precisely. Agents mirror the direction of the human effect (open-ended elicitation increases diversity in both cohorts), but their absolute diversity is consistently lower, and on Logo, Interior, and Food the gap is roughly a factor of two (e.g., Food, open-ended: humans .21 vs. agents .14). This bounds

Table 3: Persona agents produce higher visual diversity than generic baselines across all tasks. Values show CLIP Distance with Dispersion in parentheses. Δ is computed from the two-decimal values shown.

Task	Generic	Persona	$\Delta(\%)$
UI Design	.22 (.11)	.29 (.12)	+32% (+9%)
Packaging	.18 (.09)	.24 (.10)	+33% (+11%)
Logo	.08 (.03)	.12 (.05)	+50% (+67%)
Interior	.05 (.02)	.14 (.06)	+180% (+200%)
Food	.07 (.03)	.14 (.06)	+100% (+100%)
Avg	.12 (.06)	.19 (.08)	+58% (+33%)

the “virtual focus group” framing in two ways. First, it bounds the claim: FOCUSGEN expands exploration relative to a designer’s solo prompting and relative to a generic assistant (Evaluation 1), but it does not reproduce the variability that a comparable human cohort would supply. A designer who treats the agent gallery as a census of audience reactions would be systematically under-exposed to the tails of real preference distributions. Second, it suggests a mechanism. LLM outputs are known to homogenize across users and prompts [15, 23], and all of our agents share a single base model; demographic profiles and procedural backstories shift that model’s outputs but evidently do not recover the idiosyncrasy of real individuals. Consistent with this account, the gap is smallest on tasks with strong conventional structure (UI design, packaging—where agents approach human diversity) and largest where personal taste dominates (logos, interiors, food). The practical implication is that FOCUSGEN should be used as a divergence scaffold—a fast first pass that surfaces directions worth testing with real audiences—rather than as a substitute for audience research, and we have calibrated our claims throughout the paper accordingly. Closing the gap, for example via heterogeneous base models or personas grounded in real interview data [36], is a primary direction for future work.

Limitations of This Comparison. We acknowledge that the Structured and Open-Ended conditions differ in both question *format* and *degree of constraint* (five prescribed dimensions vs. agent-defined dimensions), so we cannot isolate which factor drives the diversity gains. Nonetheless, the finding provides actionable guidance: when the goal is divergent exploration, open-ended questions that let agents self-define relevant dimensions produce more diverse outputs.

5 User Study: FOCUSGEN in Professional Workflows

We conducted an exploratory qualitative study with 16 creative professionals to understand how the persona-driven, parallel-divergent workflow functions in professional ideation practice. Each session comprised two parts: a grounding task in which participants worked on a brief with their own preferred generative AI tool, followed by the same brief with FOCUSGEN (populated with U.S. census demographics [9] and procedural backstories). The first session was designed to ground participants in a concrete instance of their normal workflow that they could articulate against—not to serve

as a controlled comparison condition (see Study Design Rationale below). We designed the study to answer:

- RQ1:** How do creative professionals experience and evaluate agent-driven parallel ideation—what value and what limitations do they perceive in it within their workflows?
- RQ2:** In what ways do creative professionals utilize the diverse agents and their associated preferences to explore a design space and generate novel ideas?

5.1 Participants

We recruited 16 creative professionals (13 female, 3 male; avg. 7.75 years of experience) through online ads and personal contacts. All had professional experience with generative AI tools. The most commonly used tools were ChatGPT and Midjourney. A detailed summary is in Table 5 in the Appendix.

5.2 Study Task

Participants chose one of six creative ideation tasks (e.g., logo design for a fintech company, poster for a coffee festival; see Table 6 in the Appendix). They were asked to produce 3-5 distinct visual concepts with reference images and rationales, mimicking early-stage creative workflows.

5.3 Procedure

The study was conducted remotely via Zoom (~2 hours per session, audio/video recorded with IRB approval). Participants completed two tasks in fixed order:

Task 1: Baseline. Participants selected a task adjacent to (but distinct from) their core expertise and performed ideation using their preferred generative AI tool, compiling 3-5 concepts in a shared Google Doc.

Task 2: FOCUSGEN. After a 15-minute tutorial framing the system as a “virtual focus group,” participants performed the same task using FOCUSGEN and compiled their findings.

Study Design Rationale. We used a fixed order (familiar tool first, FOCUSGEN second) so that participants could anchor their reflections in a fresh, concrete experience of their established workflow. This design choice, together with the differing underlying models (participants’ tools vs. Gemini/Imagen) and the difference in output volume (one image per generation vs. 20–30 in parallel), makes between-condition comparisons uninterpretable, and we do not make them. The Likert battery in Figure 3 is reported as a description of participants’ absolute experience with FOCUSGEN. Two items require additional caution and should be read as exploratory. Q11 is comparative by construction (“...more effective than iterating on text prompts”) and inherits all of the confounds above; its near-neutral mean ($M = 3.9$) should not be read as evidence in either direction. Q7 (“The agents suggested ideas that I hadn’t thought of”) closely mirrors the system’s premise and is therefore vulnerable to demand characteristics; in our analysis we weight the behavioral and artifact evidence—the specific concepts participants adopted, and the follow-up interview questions they formulated—more heavily than this self-report. Comparative claims

would require a counterbalanced or between-subjects design with a matched generation backend, which we leave to future work.

Analysis. Two researchers independently coded the interview transcripts and submitted concepts using an open coding methodology, then collaboratively grouped codes into the themes presented below.

5.4 Results

Our analysis of the sessions and interviews revealed how creative professionals experienced and utilized the persona-driven workflow. We present our findings thematically, supported by qualitative feedback and descriptive statistics about the participants’ experience with FOCUSGEN. Full survey results are shown in Figure 3. We emphasize that the Likert items assess participants’ experience with FOCUSGEN in absolute terms; due to the confounds described above, we do not compare scores across conditions.

5.4.1 FOCUSGEN Unifies Visual Ideation in a Single Workflow. Addressing **RQ1**, our first finding relates to how participants experienced the FOCUSGEN workflow. When reflecting on their general practice, participants described their typical ideation process as fractured—many avoided generative tools altogether, preferring to browse visual-first sites like Pinterest. Those who did use AI described a cumbersome, two-stage process: first using an LLM for text ideas, then translating those into image prompts.

Participants found FOCUSGEN’s single-workflow approach appealing after a brief adjustment period. While initially unfamiliar with asking preference questions instead of issuing direct commands, they quickly adapted. They appreciated receiving 20–30 unique visual concepts at once—a property of the system’s parallel agent architecture—which mirrored the experience of browsing a mood board. As participant P12 noted, this meant they received “a whole board of visual starting points at once.” Participant P5 similarly emphasized that “by setting just a broad direction at the beginning, I could receive results generated from various perspectives all at once.”

5.4.2 FOCUSGEN Builds Confidence by Simulating Audience Feedback. Further answering **RQ1**, we found that FOCUSGEN builds designers’ confidence, especially in unfamiliar domains (Q10: $M=5.7$, $SD=1.1$). Many participants used their second session with FOCUSGEN to validate the ideas they had developed in the first. For example, participant P1, a brand designer working on a clothing task, used the platform to check if his initial concepts aligned with the preferences of a simulated target audience. After seeing agents independently generate concepts similar to his own initial ideas (Figure 5 in the Appendix), participant P1 felt significantly more confident and used that confidence as a basis for further, more adventurous exploration. In another example, P10, during the clothing task, asked step-by-step questions about text placement, colors, materials, and fit, receiving consistent preferences from AI agents (dark tones, long sleeves) along with specific suggestions (detachable hood, additional pockets). Participant P10 described this as “a good way for designers to understand how people think without directly conducting user research,” emphasizing the value of rapid validation. In these cases, FOCUSGEN served as a “sanity check,” allowing designers to quickly test their intuition against a simulated audience.

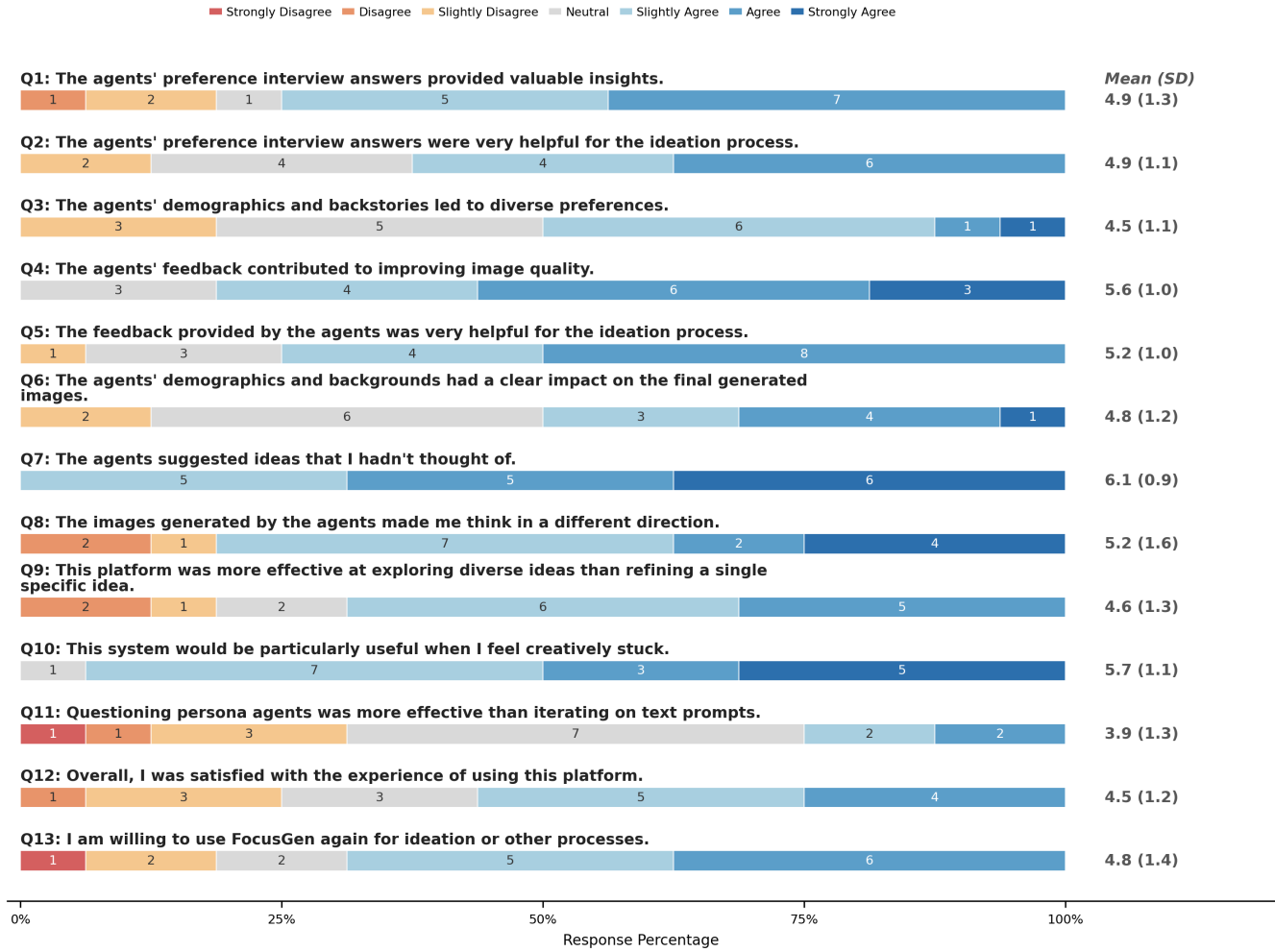


Figure 3: Overall usability survey questionnaires and results for our system FOCUSGEN. (Q7 and Q11 are exploratory items; see Section 5.3.)

5.4.3 Personas Generate Novel, Contextually-Rich Ideas. Our central finding, speaking to both **RQ1** and **RQ2**, is how personas act as a catalyst for novel ideas that push users in new directions by providing diverse perspectives (Q7: $M=6.1$, $SD=0.9$; Q9: $M=4.6$, $SD=1.3$). A clear example was the “Coffee Brew 2025” coffee festival poster task. In the baseline condition, participants used web searches or direct prompting, resulting in more conventional designs. With FOCUSGEN, the results were far more diverse and contextually-aware (as seen in Figure 4 in the Appendix). Participant P6 was particularly impressed by a poster featuring a caffeine molecule—a completely different approach that she had not thought of. Participant P9 highlighted a concept where a drop of coffee fell onto a famous city landmark, an idea she found unique and clever.

Crucially, this novelty was not just visual; it was contextual. Participants reported that the agents’ preference interview answers provided valuable insights (Q1: $M=4.9$, $SD=1.3$; Q2: $M=4.9$, $SD=1.1$).

Unlike static images from Pinterest, these textual responses explained the “why” behind the “what,” serving a role similar to a docent explaining the intent behind a piece of art. Participant P11 described it: “For Assignment 2 [FOCUSGEN], I felt like I engaged more in critical thinking, trying to understand the rationale behind the designs that the agents generated.” Participant P16 similarly noted that interviewing agents enabled deeper reflection and thinking compared to simple prompting for image generation. This contextual understanding was generative in itself, often prompting participants to formulate new interview questions to explore these unexpected creative directions further.

5.4.4 Demographics as a Creative Lever for Conceptual Exploration. Providing an answer to **RQ2**, our study revealed how participants used demographic filters as a creative lever to probe the design space in unexpected ways. While the perceived impact of demographics was moderately significant overall (Q6: $M=4.8$, $SD=1.2$), certain tasks provided salient examples of their influence.

The most striking instance occurred during the “Beauty Expo 2025” mascot task. Participants P9 and P12 initially sampled agents from one demographic. When they broadened their exploration to include another, the results suddenly included robot characters—a concept that had not appeared before (see Figure 6 in the Appendix). Participant P12 found this revelatory: *“It was really interesting... when I sampled agents that were [Gender B], some characters came out to be robots, which never appeared when I set the agents to all [Gender A].”* This discovery illustrates both the potential and the risk of demographic conditioning. On one hand, changing a single filter surfaced an entirely new conceptual direction that no participant had considered—exactly the kind of “blind spot” FOCUSGEN is designed to reveal. On the other hand, the association between a specific demographic and a specific concept (robots in a beauty context) may reflect stereotypical patterns in the underlying model rather than genuine audience preferences. This tension is inherent to persona-based systems and underscores the importance of the designer’s curatorial role: the value lies not in accepting any single agent’s output as representative, but in using the diversity of outputs as provocations that expand the design space. We discuss mitigation strategies in Section 6.1.

This principle of contextual relevance extended beyond gender to agents’ professional backgrounds. For the fintech logo task, participant P16 noted how a healthcare worker agent logically introduced medical-inspired design elements. Similarly, participant P9 found that the novel idea of using a caffeine molecule for a coffee poster *“made sense”* because it was generated by an agent with a background in engineering. These instances show how an agent’s specific life experience can generate logical yet unexpected conceptual blends that participants found valuable.

More broadly, participants valued demographic filters for grounding ideation in specific audience contexts, though this capability also raises questions about reinforcing stereotypes that we address in Section 6.1.

5.4.5 Participants Positioned FOCUSGEN as a Tool for Divergent Exploration. Finally, answering **RQ2**, participants quickly formed a strong mental model of FOCUSGEN’s ideal role. They agreed it would be most useful when creatively stuck (Q10: $M=5.7$, $SD=1.1$), and they leaned toward viewing it as better suited to exploring diverse ideas than to refining a single concept (Q9: $M=4.6$, $SD=1.3$; see also Q7: $M=6.1$, $SD=0.9$). Participant P6 evaluated FOCUSGEN as *“suitable for ideation while finding diverse references,”* emphasizing its utility for idea divergence. Participant P10 noted that *“I could re-examine elements I took for granted and break existing preconceptions,”* recognizing FOCUSGEN’s value as a tool for providing new perspectives. This positions FOCUSGEN as a complementary, upstream tool whose primary value lies in breaking creative fixation at the earliest stages of a project.

6 Discussion

In this work, we introduced FOCUSGEN, an interactive system that brings external audience perspectives into visual design exploration through a virtual focus group of persona agents. Our evaluations provide converging evidence that persona conditioning expands the diversity of generated design alternatives beyond what a generic

assistant produces, and our qualitative study illustrates how creative professionals put that expanded spectrum to work—while also delineating what we have not shown: that agent outputs are discriminably faithful to their specific personas, or that they represent the preferences of real demographic groups (Sections 4 and 6.4).

6.1 Ethical Considerations: Stereotyping and the Appeal of the Characterful

The demographic grounding that enables FOCUSGEN to simulate diversity also risks reproducing demographic stereotypes [28, 51]. Narrative backstories and preference interviews shift agents from demographic shorthand toward individualized traits [36], but unlike Park et al.’s agents, which are grounded in interviews with real people, our backstories are themselves LLM-generated from demographic seeds. The mitigation is therefore partial and unaudited: the backstory-generation process may launder stereotypical associations into fluent individual narratives, and we did not conduct a formal bias audit.

Our own findings sharpen this concern rather than soften it. In the Beauty Expo case (Section 5.4.4), changing a single gender filter produced robot mascots, and participants found the outputs delightful and adopted the direction. Designer curation—which an earlier version of this paper offered as the primary safeguard—operated here in the opposite direction: the surprisingness that makes an output feel like a discovered blind spot is also what makes a stereotyped association attractive and likely to be incorporated. In a tool built to reward surprise, novelty-seeking and stereotype-adoption can be behaviorally indistinguishable. Similarly, P10 described the system as “a good way for designers to understand how people think without directly conducting user research”—a reading our validation does not license, since we have not shown that agent preferences track those of real demographic groups.

We see three concrete obligations for systems like FOCUSGEN. First, provenance and friction: the interface already attaches each image to the generating agent’s full profile; it should additionally surface when a visual direction correlates strongly with a single demographic attribute across the cohort, prompting the designer to test that direction against other segments before adopting it. Second, framing: onboarding and documentation must present outputs as provocations for exploration, not as evidence of audience preference, since our participants did not always maintain this distinction unprompted. Third, auditing and validation: deployment beyond a research prototype requires a formal bias audit of the backstory- and preference-generation pipeline and validation of persona outputs against real demographic samples (Section 6.4). Until then, the appropriate claim for FOCUSGEN is that it diversifies a designer’s option set—not that it represents an audience.

6.2 The Designer’s Role: Agency and Ownership

By automating the generation of visual concepts from multiple persona perspectives, FOCUSGEN performs substantial creative labor with minimal real-time human input, raising questions about creative agency [18]. We position FOCUSGEN deliberately as an *upstream* ideation tool: the system generates a landscape of starting points, and the designer’s expertise enters in curating, interpreting, and building upon these seeds. Our user study supports

this—participants described the system as most valuable for early exploration, and P11 reported engaging in *more* critical thinking because interpreting diverse agent perspectives demanded active judgment. The interview question design step, where the designer’s creative intent shapes the system’s behavior, may be the key moment of *appropriation* [14]. Future work should investigate how extended use affects designers’ sense of ownership.

6.3 Implications for Creativity Support Tool Design

6.3.1 From Dyadic to Polyphonic Co-Creation. Current Human-AI interaction is predominantly dyadic. FOCUSGEN proposes a shift toward **polyphonic** co-creation, extending crowd-powered design [2, 27] into the “in silico” domain, where interfaces present *conflicting* perspectives simultaneously rather than optimizing for a single response [6].

6.3.2 Narrative-Anchored Divergence over Stochastic Variation. Our results suggest that *contextual novelty* is far more valuable than stochastic variation [11, 38]. Persona-driven deviations provide a **semantic anchor**—a narrative logic that makes novelty plausible (e.g., an engineering agent suggesting a caffeine molecule). This extends work on meaningful diversity [19]: narrative grounding transforms diversity from noise into actionable inspiration.

6.3.3 AI as a Mirror for Creative Self-Efficacy. Participants used the system not just for inspiration but for internal validation—seeing intuitions reflected by a simulated audience gave them confidence to take risks [4]. This suggests AI simulation can function as a “safe-to-fail” sandbox [42] for creative risk-taking.

6.3.4 Eliciting Dimensions versus Imposing Them. Our open-ended-versus-structured finding connects directly to the dimension–value framework of Proxona [10], in which audience personas are organized along creator-curated dimensions, and to dimensional-reasoning tools more broadly [8, 45, 46]. Structured interviews in FOCUSGEN function like designer-imposed dimensions: they guarantee coverage of axes the designer already considers relevant, which supports systematic comparison, but in our data they compress output diversity. Open-ended interviews instead let each persona nominate its own dimensions, trading comparability for breadth—and it is precisely this self-definition that appears to surface dimensions the designer had not considered. The two modes are complementary: a practical workflow might open with open-ended elicitation to map the space, then switch to structured questions to compare cohorts along the dimensions that emerged. This also clarifies our relationship to convergent persona systems such as PosterMate [40] and PersonaFlow [29]: those systems concentrate persona perspectives onto one artifact or one research question, which suits refinement; the parallel-divergent topology suits the earlier moment when the designer does not yet know which artifact to refine.

6.4 Limitations and Future Work

6.4.1 Synthetic Nature of Personas and the Missing Discriminative Test. Our synthetic agents remain models without lived experience, and two validation gaps follow. First, discriminative faithfulness: our Technical Validation shows that the refinement loop improves alignment with a given preference profile, but we did not test whether outputs are discriminably tied to their specific persona—for example, whether participants prefer images generated from their own profile over images generated from another participant’s profile, or whether persona-conditioned refinement outperforms refinement by a randomly assigned persona or a generic critic on the same profiles. This test is the right next step and is inexpensive, since it reuses the existing pipeline. Second, representativeness: we did not validate that synthetic agents’ self-generated preferences match those of the demographic groups they are sampled from, and it is unknown to what extent the underlying models have learned genuine correlations between demographics and aesthetic preference (the illustrative carpenter example in Section 3.2.2 should be read as a description of the mechanism, not as a validated association). A valuable community resource here would be a calibration dataset of real aesthetic preferences across demographic strata, against which persona pipelines could be re-validated as underlying models change; absent such a dataset, results like ours are tied to specific model versions (gemini-2.5-flash, imagen-3.0-generate) and must be re-established when models change. Future work should also ground personas in real, anonymized interview data [36] and conduct systematic bias audits of the backstory-generation process.

6.4.2 Non-Monotonic Refinement. The refinement process is non-monotonic—later iterations do not always improve upon earlier ones. Our design choice to have agents review the *entire* image sequence and select the best (rather than defaulting to the last) addresses this, but future work should investigate more principled convergence criteria.

6.4.3 Generalizability. Our study was limited to 16 creative professionals from a non-U.S. country, and our Agent Bank uses U.S. census data. This cultural mismatch may inflate perceived novelty—some concepts may have felt surprising due to cultural distance rather than the persona mechanism itself. Future work should match agent demographics to the designer’s target market, replicate with broader participant samples, diverse cultural datasets, and extend beyond static images to domains like prototyping and video.

7 Conclusion

Creative professionals design for audiences, not just themselves—yet current text-to-image exploration tools derive diversity entirely from the designer’s own input, limiting discovery to what the designer already knows to look for. In this paper, we introduced FOCUSGEN, an interactive system that brings external perspectives into visual design exploration through a virtual focus group of persona agents. Each agent, constructed from demographic data, a procedural backstory, and elicited aesthetic preferences, independently generates visual concepts conditioned on that profile—surfacing directions the designer would not have explored alone. Our evaluations showed that agent diversity translates directly into visual

diversity, and that open-ended preference interviews expand the design space by letting agents define their own aesthetic dimensions. A study with 16 creative professionals found that FOCUSGEN helped them discover unanticipated directions—from caffeine molecules on coffee posters to robots in beauty mascots—overcome fixation, and build confidence through simulated audience feedback.

These perspectives are simulations: our evaluations show that they expand the diversity of a designer’s option set and that designers find the expansion useful, while leaving open—and we have made explicit—the question of whether they are faithful to the specific people and groups they are modeled on.

This work contributes to a broader shift in human-AI creative partnership: from engineering prompts for a single AI toward curating a diverse, persona-driven “virtual focus group” that surfaces the blind spots designers need to see.

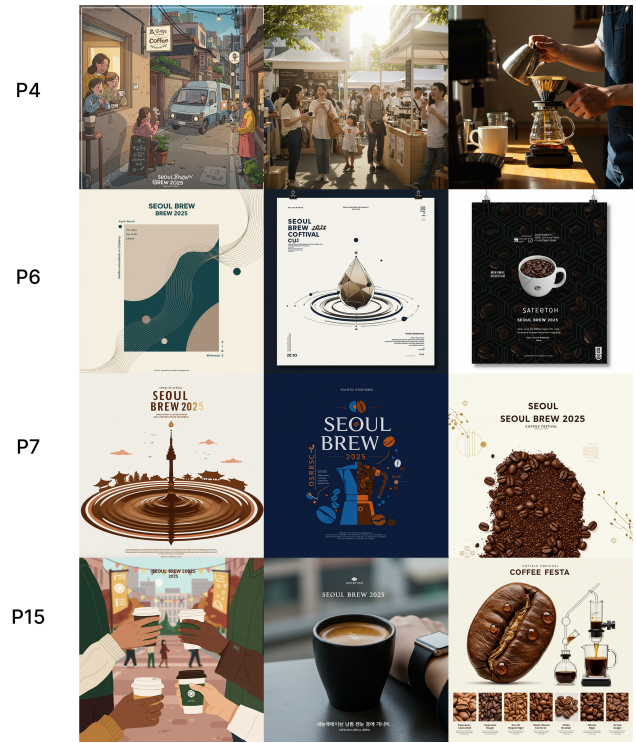
References

- [1] Maneesh Agrawala. 2023. Unpredictable Black Boxes are Terrible Interfaces. <https://magrawala.substack.com/p/unpredictable-black-boxes-are-terrible>
- [2] Salvatore Andolina, Hendrik Schneider, Joel Chan, Khalil Klouche, Giulio Jacucci, and Steven Dow. 2017. Crowdboard: Augmenting In-Person Idea Generation with Real-Time Crowds. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition (C&C '17)*. Association for Computing Machinery, New York, NY, USA, 106–118. doi:10.1145/3059454.3059477
- [3] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* 31, 3 (July 2023), 337–351. doi:10.1017/pan.2023.2 arXiv:2209.06899 [cs].
- [4] Albert Bandura. 1977. Self-efficacy: Toward a Unifying Theory of Behavioral Change. *Psychological Review* 84, 2 (1977), 191–215. doi:10.1037/0033-295X.84.2.191
- [5] Cynthia L. Bennett and Daniela K. Rosner. 2019. The Promise of Empathy: Design, Disability, and Knowing the “Other”. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, 1–13. doi:10.1145/3290605.3300528
- [6] Susanne Bødker. 2006. When Second Wave HCI Meets Third Wave Challenges. In *Proceedings of the 4th Nordic Conference on Human-Computer Interaction (NordiCHI '06)*. ACM, 1–8. doi:10.1145/1182475.1182476
- [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. doi:10.48550/arXiv:2304.09337 arXiv:2304.09337 [cs].
- [8] Alice Cai, Steven R Rick, Jennifer L Heyman, Yanxia Zhang, Alexandre Filipowicz, Matthew Hong, Matt Klenk, and Thomas Malone. 2023. DesignAID: Using Generative AI and Semantic Diversity for Design Inspiration. In *Proceedings of The ACM Collective Intelligence Conference*. ACM, Delft Netherlands, 1–11. doi:10.1145/3582269.3615596
- [9] Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. 2024. PERSONA: A Reproducible Testbed for Pluralistic Alignment. doi:10.48550/arXiv:2407.17387 arXiv:2407.17387 [cs].
- [10] Yoonseo Choi, Eun Jeong Kang, Seulgi Choi, Min Kyung Lee, and Juho Kim. 2025. Proxona: Supporting Creators’ Sensemaking and Ideation with LLM-Powered Audience Personas. doi:10.48550/arXiv:2408.10937 arXiv:2408.10937 [cs].
- [11] Gabriele Corso, Yilun Xu, Valentin de Bortoli, Regina Barzilay, and Tommi Jaakkola. 2023. Particle Guidance: non-I.I.D. Diverse Sampling with Diffusion Models. doi:10.48550/arXiv:2310.13102 arXiv:2310.13102 [cs].
- [12] Nathan Crilly and Carlos Cardoso. 2017. Where Next for Research on Fixation, Inspiration and Creativity in Design? *Design Studies* 50, 1–38. doi:10.1016/j.destud.2017.02.001
- [13] Nigel Cross. 2004. Expertise in design: an overview. *Design Studies* 25, 5 (Sept. 2004), 427–441. doi:10.1016/j.destud.2004.06.002
- [14] Alan Dix. 2007. Designing for Appropriation. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers (BCS-HCI '07)*. British Computer Society, 27–30.
- [15] Anil R. Doshi and Oliver P. Hauser. 2024. Generative AI Enhances Individual Creativity but Reduces the Collective Diversity of Novel Content. *Science Advances* 10, 28 (2024), eadn5290. doi:10.1126/sciadv.adn5290
- [16] Steven P. Dow, Alana Glasco, Jonathan Kass, Melissa Schwarz, Daniel L. Schwartz, and Scott R. Klemmer. 2010. Parallel Prototyping Leads to Better Design Results, More Divergence, and Increased Self-Efficacy. *ACM Transactions on Computer-Human Interaction* 17, 4 (2010), 1–24. doi:10.1145/1879831.1879836
- [17] Yingchaojie Feng, Xingbo Wang, Kam Kwai Wong, Sijia Wang, Yuhong Lu, Minfeng Zhu, Baicheng Wang, and Wei Chen. 2023. PromptMagician: Interactive Prompt Engineering for Text-to-Image Creation. *IEEE Transactions on Visualization and Computer Graphics* (2023), 1–11. doi:10.1109/TVCG.2023.3327168 arXiv:2307.09036 [cs].
- [18] Jonas Frich, Lindsay MacDonald Vermeulen, Christian Remy, Michael Mose Biskjaer, and Peter Dalsgaard. 2019. Mapping the Landscape of Creativity Support Tools in HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–18. doi:10.1145/3290605.3300619
- [19] 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. doi:10.48550/arXiv.2504.13392 arXiv:2504.13392 [cs].
- [20] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2023. Optimizing Prompts for Text-to-Image Generation. doi:10.48550/arXiv.2212.09611 arXiv:2212.09611 [cs].
- [21] Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkill™: a Bayesian skill rating system. In *Proceedings of the 20th International Conference on Neural Information Processing Systems (NIPS'06)*. MIT Press, Cambridge, MA, USA, 569–576.
- [22] Yu-Kai Hung, Yun-Chien Huang, Ting-Yu Su, Yen-Ting Lin, Lung-Pan Cheng, Bryan Wang, and Shao-Hua Sun. 2024. SimTube: Generating Simulated Video Comments through Multimodal AI and User Personas. doi:10.48550/arXiv.2411.09577 arXiv:2411.09577 [cs].
- [23] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users’ Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM. doi:10.1145/3544548.3581196
- [24] David G. Jansson and Steven M. Smith. 1991. Design fixation. *Design Studies* 12, 1 (Jan. 1991), 3–11. doi:10.1016/0142-694X(91)90003-F
- [25] Morten B. Jensen, Christer W. Elverum, and Martin Steinert. 2017. Eliciting Unknown Unknowns with Prototypes: Introducing Prototrials and Prototrial-Driven Cultures. *Design Studies* 49 (2017), 1–31. doi:10.1016/j.destud.2016.12.002
- [26] Merlijn Kouprie and Froukje Sleswijk Visser. 2009. A Framework for Empathy in Design: Stepping Into and Out of the User’s Life. *Journal of Engineering Design* 20, 5 (2009), 437–448. doi:10.1080/09544820902875033
- [27] Yuki Koyama, Daisuke Sakamoto, and Takeo Igarashi. 2014. Crowd-powered parameter analysis for visual design exploration. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. ACM, Honolulu Hawaii USA, 65–74. doi:10.1145/2642918.2647386
- [28] Yuxuan Li, Hirokazu Shirado, and Sauvik Das. 2025. Actions Speak Louder than Words: Agent Decisions Reveal Implicit Biases in Language Models. doi:10.48550/arXiv.2501.17420 arXiv:2501.17420 [cs].
- [29] Yiren Liu, Pranav Sharma, Mehak Oswal, Haijun Xia, and Yun Huang. 2025. PersonaFlow: Designing LLM-Simulated Expert Perspectives for Enhanced Research Ideation. In *Proceedings of the 2025 ACM Conference on Designing Interactive Systems (DIS '25)*. ACM.
- [30] Kurt Luther, Jari-Lee Tolentino, Wei Wu, Amy Pavel, Brian P. Bailey, Maneesh Agrawala, Björn Hartmann, and Steven P. Dow. 2015. Structuring, Aggregating, and Evaluating Crowdsourced Design Critique. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 473–485. doi:10.1145/2675133.2675283
- [31] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative Refinement with Self-Feedback. doi:10.48550/arXiv.2303.17651 arXiv:2303.17651 [cs].
- [32] Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedar-madji, Eran Kohen Behar, and David M. Chan. 2024. Virtual Personas for Language Models via an Anthology of Backstories. doi:10.48550/arXiv.2407.06576 arXiv:2407.06576 [cs].
- [33] Jonas Oppenlaender. 2022. The Creativity of Text-to-Image Generation. In *Proceedings of the 25th International Academic Mindtrek Conference*. 192–202. doi:10.1145/3569219.3569352 arXiv:2206.02904 [cs].
- [34] Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. doi:10.48550/arXiv.2304.03442 arXiv:2304.03442 [cs].
- [35] Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. doi:10.48550/arXiv.2208.04024 arXiv:2208.04024 [cs].
- [36] Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative Agent Simulations of 1,000 People. doi:10.48550/arXiv.2411.10109 arXiv:2411.10109 [cs].

- [37] Michael J. Ryan, Omar Shaikh, Aditri Bhagirath, Daniel Frees, William Held, and Diyi Yang. 2025. SynthesizeMe! Inducing Persona-Guided Prompts for Personalized Reward Models in LLMs. doi:10.48550/arXiv.2506.05598 arXiv:2506.05598 [cs].
- [38] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Roman M. Weber. 2024. CADs: Unleashing the Diversity of Diffusion Models through Condition-Annealed Sampling. doi:10.48550/arXiv.2310.17347 arXiv:2310.17347 [cs] version: 4.
- [39] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A Trainable Agent for Role-Playing. doi:10.48550/arXiv.2310.10158 arXiv:2310.10158 [cs].
- [40] Donghoon Shin, Daye Lee, Gary Hsieh, and Gilbert Yat-Yin Chan. 2025. Poster-Mate: Audience-driven Collaborative Persona Agents for Poster Design. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. ACM. doi:10.1145/3746059.3747769
- [41] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. doi:10.48550/arXiv.2303.11366 arXiv:2303.11366 [cs].
- [42] David J. Snowden and Mary E. Boone. 2007. A Leader's Framework for Decision Making. *Harvard Business Review* 85, 11 (2007), 68–76.
- [43] 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*. ACM, Honolulu HI USA, 1–19. doi:10.1145/3613904.3642847
- [44] Hariharan Subramonyam, Roy Pea, Christopher Lawrence Pondoc, Maneesh Agrawala, and Colleen Seifert. 2024. Bridging the Gulf of Envisioning: Cognitive Design Challenges in LLM Interfaces. doi:10.48550/arXiv.2309.14459 arXiv:2309.14459 [cs].
- [45] 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*. 1–26. doi:10.1145/3613904.3642400 arXiv:2310.12953 [cs].
- [46] 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*. ACM, Yokohama Japan, 1–26. doi:10.1145/3706598.3714211
- [47] Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization. doi:10.48550/arXiv.2406.01171 arXiv:2406.01171 [cs].
- [48] Kavana Venkatesh, Connor Dunlop, and Pinar Yanardag. 2025. CREA: A Collaborative Multi-Agent Framework for Creative Image Editing and Generation. In *Advances in Neural Information Processing Systems*, D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (Eds.), Vol. 38. Curran Associates, Inc., 171332–171392. https://proceedings.neurips.cc/paper_files/paper/2025/file/fa41e9d5dfcc97cd9eed99f001aa28e5-Paper-Conference.pdf
- [49] Samitha Wadinambiarachchi, Ryan Kelly, Shreyas Pareek, Qiushi Zhou, and Eduardo Velloso. 2024. The Effects of Generative AI on Design Fixation and Divergent Thinking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM. doi:10.1145/3613904.3642919
- [50] Zhijie Wang, Yuheng Huang, Da Song, Lei Ma, and Tianyi Zhang. 2024. PromptCharm: Text-to-Image Generation through Multi-modal Prompting and Refinement. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21. doi:10.1145/3613904.3642803 arXiv:2403.04014 [cs].
- [51] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *2022 ACM Conference on Fairness Accountability and Transparency*. ACM, Seoul Republic of Korea, 214–229. doi:10.1145/3531146.3533088
- [52] Yutong Xie, Zhaoying Pan, Jinge Ma, Luo Jie, and Qiaozhu Mei. 2023. A Prompt Log Analysis of Text-to-Image Generation Systems. In *Proceedings of the ACM Web Conference 2023*. 3892–3902. doi:10.1145/3543507.3587430 arXiv:2303.04587 [cs].
- [53] Anbang Xu, Shih-Wen Huang, and Brian P. Bailey. 2014. Voyant: Generating Structured Feedback on Visual Designs Using a Crowd of Non-Experts. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. ACM, 1433–1444. doi:10.1145/2531602.2531604
- [54] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–21. doi:10.1145/3544548.3581388



(a) Visual concept from a baseline tool.

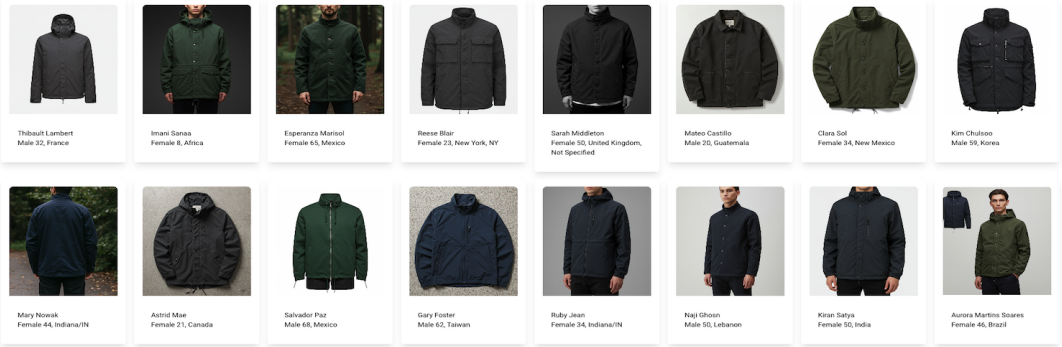


(b) Visual concept from FOCUSGEN.

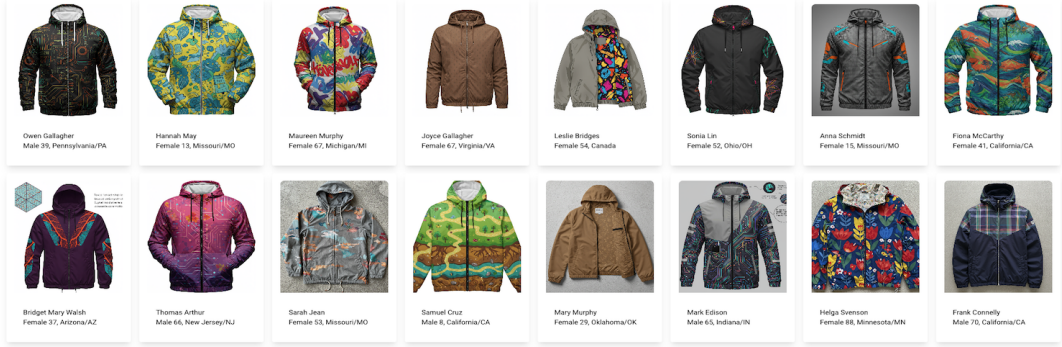
Figure 4: Comparison of concepts generated for the “Coffee Brew 2025” poster task. Concepts from baseline tools (left) were often conventional, while FOCUSGEN (right) generated more diverse and contextually-aware ideas, such as the caffeine molecule and the coffee drop on the city landmark, which participants found novel.



(a) Images generated by Participant P1 using his preferred AI image tool.



(b) First batch of images generated by Participant P1 using FocusGEN.



(c) Subsequent batch of images generated by Participant P1 using FocusGEN.

Figure 5: An example of how participant P1 used FocusGEN to build creative confidence: P1’s initial concepts, generated with a baseline tool as shown in 5a, were validated when FocusGEN’s agents independently produced similar ideas as shown in 5b. This confirmation gave participant P1 the confidence to then use the system to explore more adventurous, novel concepts (in 5c).

Input Prompt Type	Prompt Content	Token Count
Demographic Profile	• Name: Mia • Age: 29 • Gender: Female • Race: White alone • Education: graduate degree • ... • Place of Birth: New Mexico/NM	≈ 150
Backstory	I was born 29 years ago in New Mexico, a state that always felt vast and quiet, painted with colors of red earth and infinite sky. My childhood wasn't particularly eventful, at least not in the dramatic sense. It was... Growing up in New Mexico was a calm existence. We... My focus was always on my studies. High school was a period of intense academic application. I thrived on... When it came time for college, I moved away from New Mexico, drawn by a program that offered a rigorous academic curriculum. I pursued a Bachelor's degree, followed quickly by a Master's degree in business, with a specialization in risk management and finance... It was during my Master's program that I met my husband, Alex... After graduating, I went straight into the insurance industry... I spend my days sifting through information, scrutinizing facts, determining coverage and liability...	≈ 1000
Preference Profile	Here are your interview responses for an ideal bedroom: • I would use matte white paint for the walls with one accent wall in muted sage green. The floor would... • I want a low-profile platform bed with a walnut wood frame. The bedding would be... • I'd include a small desk in the corner near the window, a full-length mirror...	≈ 400
Instruction Prompt	Your task is to roleplay as this character and answer questions based on your profile. Your responses should reflect your character's personality, background, and experiences.	≈ 200

Table 4: The prompt for a user-simulating persona agent is constructed by concatenating the user's demographic profile and personal backstory, their elicited preference interview data, and a final instruction for the agent to adopt this profile.

ID (Gender, Age)	Profession/Major	YoE	Industry/Major (Employment)	Preferred Gen AI Tools*
P1 (M, 30s)	Brand Designer	8	Finance (Full-time)	Adobe AI , Shutterstock AI
P2 (F, 20s)	Art Director	3	Ad Agency H (Full-time)	ChatGPT , Midjourney
P3 (F, 30s)	Art Director	10	Ad Agency H (Full-time)	ChatGPT
P4 (F, 30s)	Art Director	10	Ad Agency D (Full-time)	ChatGPT , Midjourney
P5 (F, 30s)	Brand Designer	11	Ad Agency C (Full-time)	ChatGPT , Adobe AI, Midjourney
P6 (F, 30s)	Art Director	10	Ad Agency H (Full-time)	ChatGPT , Midjourney, Adobe AI
P7 (F, 30s)	Art Director	8	Ad Agency H (Full-time)	ChatGPT , Gemini
P8 (F, 30s)	Fashion(Jewelry) Designer	3	Startup (Full-time)	ChatGPT , Midjourney
P9 (F, 30s)	Art Director	10	Ad Agency H (Full-time)	ChatGPT
P10 (F, 20s)	Master’s Student	4	Industrial Design (Full-time)	ChatGPT , Midjourney
P11 (M, 20s)	Master’s Student	3	Industrial Design (Full-time)	ChatGPT , Midjourney
P12 (F, 20s)	Freelancer	4	Self-Employed (Part-time)	ChatGPT
P13 (F, 30s)	Art Director	8	Ad Agency H (Full-time)	ChatGPT
P14 (F, 30s)	Freelancer	10	Self-Employed (Part-time)	ChatGPT , Midjourney
P15 (F, 30s)	Art Director	8	Ad Agency H (Full-time)	ChatGPT , Veo3
P16 (M, 20s)	UX Designer	4	Education Tech (Full-time)	ChatGPT , Bolt.new, DALL-E

Table 5: Summary of the 16 creative professionals who participated in our user study. The table details their demographic and professional backgrounds, years of experience, and preferred generative AI tools. The tool indicated in bold is the one each participant chose to use for the baseline condition in our study.

Table 6: An overview of the six design ideation tasks used in our user study. The tasks were designed to cover a diverse range of creative scenarios, each with a unique brief and set of constraints.

No.	Task Name	Task Overview
1	Logo Design Ideation	Collect logo references for a new fintech service, “ Future Pay ”. The core values are “ Transparency, Speed, and Trust ”.
2	Character Design Ideation	Collect references for the mascot of the “ Beauty Expo 2025 ”. The character will be used to expand the brand’s tone on packaging, app icons, and social media.
3	Clothing/Shoe Design Ideation	Collect outerwear references for the “ Urban Run 2025 ” sports festival’s capsule collection. The main item is a unisex hybrid jacket for both urban commuting and outdoor activities.
4	Ad Poster Design Ideation	Collect poster references for the “ Coffee Brew 2025 ” coffee festival. The poster should emphasize “ simplicity + a stylish lifestyle ” and will be used in cafes, at bus stops, and on social media.
5	Social Media Campaign Ideation	Collect visual references for the “ Summer Spark 2025 ” social media campaign. Ads will run on Instagram and TikTok, needing a mood that evokes “ coolness, trendiness, and a desire to share ”.
6	Freestyle Design Ideation	Assume you receive an “ open brief ” from a new client entering a new market. The client wants to visually explore their brand’s tone and mood.

Participant Ideation Images (Character Design)



Figure 6: A salient example of demographic filters influencing conceptual direction from the “Beauty Expo 2025” mascot task. Left: Images from agents sampled from Gender A only. Right: Images from agents sampled from all genders.

Question 3: Which set of images do you think is more diverse?



Figure 7: The interface for the comparative study in our diversity evaluation. Participants were shown two 3x3 grids of images for a given creative task and were asked to select which grid exhibited greater visual diversity, or to choose a tie. For each pair, one grid was generated using the Persona-Disabled (Baseline) condition, while the other was generated using the full, Persona-Enabled FOCUSGEN system.