Adaptive Prompt Elicitation for Text-to-Image Generation

Xinyi Wen
Aalto University & University of
Helsinki & ELLIS Institute Finland
Helsinki, Finland
xinyi.wen@aalto.fi

Lena Hegemann
Aalto University
Helsinki, Finland
lena.hegemann@aalto.fi

Xiaofu Jin
Aalto University
Helsinki, Finland
xjinao@connect.ust.hk

Shuai Ma
Aalto University
Helsinki, Finland
mashuai@iscas.ac.cn

Antti Oulasvirta
Aalto University & ELLIS Institute
Finland
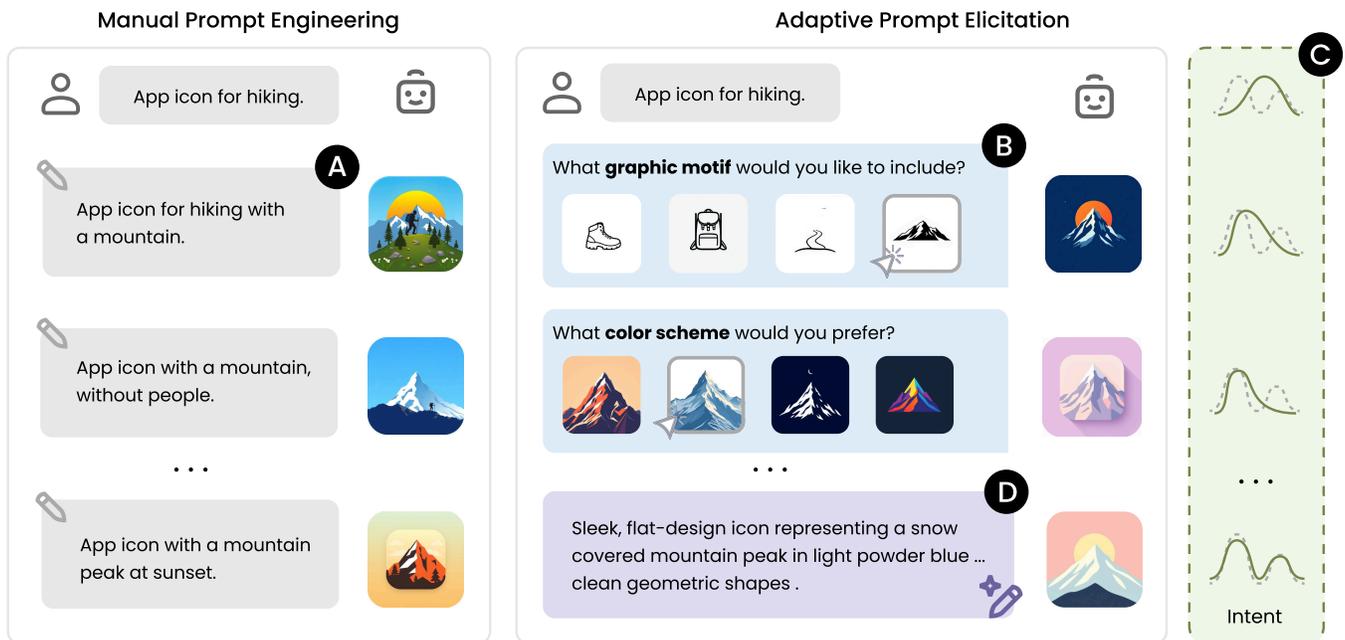Helsinki, Finland
antti.oulasvirta@aalto.fi

Figure 1: Manual prompt engineering (A) requires users to iterate extensively. ADAPTIVE PROMPT ELICITATION (APE) addresses this by (B) presenting visual queries, (C) updating its belief about user intent (solid curve) to align with the user's latent intent (dotted curve) based on responses, and (D) generating optimized prompts from its beliefs.

Abstract

Aligning text-to-image generation with user intent remains challenging, as users frequently provide ambiguous inputs and struggle with model idiosyncrasies. We propose ADAPTIVE PROMPT ELICITATION (APE), a technique that adaptively poses visual queries to help users refine prompts without extensive writing. Our technical contribution is a formulation of interactive intent inference under an information-theoretic framework. APE represents latent user intent as interpretable feature requirements using language model priors, adaptively generates visual queries, and compiles elicited requirements into effective prompts. Evaluation on IDEA-Bench and DesignBench shows that APE achieves stronger alignment with improved efficiency. A user study with 128 participants on user-defined tasks demonstrates 19.8% higher perceived alignment without increased workload. Our work contributes a principled approach to prompting that offers an effective and efficient complement to the prevailing prompt-based interaction paradigm with text-to-image models.

CCS Concepts

• **Human-centered computing** → **Collaborative interaction**; **User models**; • **Computing methodologies** → **Probabilistic reasoning**.

Keywords

Text-to-Image Generation, Prompt Optimization, Human-AI Alignment, Active Preference Learning, Information Theory

1 Introduction

Text-to-image generation models have achieved remarkable advances in synthesizing imagery from natural language descriptions [67, 68, 79], however, *aligning model outputs with user intent* remains an outstanding challenge. This stems from the difficulty of articulating rich visual concepts precisely through text [54, 62, 69] besides inherent limitations in model capacity. This problem is particularly pronounced for users who lack familiarity with prompt engineering conventions and the characteristics of the specific model [51, 87, 90]. Even experienced users struggle to anticipate how specific linguistic choices influence generation outcomes, leading to iterative trial-and-error cycles that are time-consuming and cognitively taxing [59, 85].

Our work converges on two lines of prior research. One line aims to guide users to write better prompts, via keyword-rich descriptions, style modifiers, artist references, or quality tokens [51, 59]. These approaches, however, require familiarity with technical vocabulary or interaction styles that may not be accessible to non-expert users [90]. A second line focuses on developing automatic prompt optimization methods that translate user input into high-quality prompts without requiring manual refinement. These methods implicitly assume that user intentions are clearly articulated, may leave users to refine the prompt iteratively through trial and error [54, 66].

In this paper, we introduce ADAPTIVE PROMPT ELICITATION (APE), a computational method that inverts the traditional prompting paradigm by actively eliciting user's latent intent and compiling it into effective prompts. From the user's perspective, APE acts like a human interviewer experienced in the domain of interest and asks guiding questions that direct the conversation: "Would you like this or that more?". As shown in Figure 1, after receiving an initial prompt that sets the theme for interaction, APE presents visual queries that help users progressively clarify their intent. These queries require only visual judgments rather than linguistic articulation, reducing reliance on domain knowledge and linguistic ability. Users respond through simple selections, and APE synthesizes these responses into refined prompts that capture their clarified intent. APE is designed for scenarios where users have latent visual preferences but struggle to articulate them precisely through text, a documented challenge in text-to-image generative interaction [28, 54].

Technically, to bridge the gap between user intent and the model, we establish an explicit set of interpretable visual features as a partial observation of users' latent intent, operationalized through language model priors that map natural language to structured feature requirements. APE maintains a belief distribution over possible intents and selects visual queries grounded in an information-theoretic framework [9, 64], providing a powerful model-based framework for choosing designs automatically. Specifically, it ensures that each query maximally reduces misalignment about intent while respecting constraints on user interaction time. By grounding the elicitation process in a coherent probabilistic framework, APE adaptively tailors queries to individual users and tasks, achieving well-aligned images within a few iterations.

Our complementary evaluations provide convergent evidence for APE's effectiveness and efficiency. On two text-to-image generation benchmarks, IDEA-Bench [47] and DesignBench [27], APE achieves stronger alignment between generated images and ground-truth intents compared to baseline prompting approaches, while requiring significantly fewer interaction steps. A user study with 128 participants with varying experience levels further validates these findings in ecologically valid settings with self-defined creative tasks. Participants reported a 19.8% increase in perceived alignment between generated outputs and their intention compared to manual prompting. Qualitative analysis reveals that APE's interactive elicitation process helps users discover and articulate preferences they could not initially verbalize, suggesting that the approach not only elicits existing intent but facilitates intent formation through guided exploration [52, 66].

Our primary contributions are as follows:

- We present a novel adaptive prompt elicitation method that formalizes prompt optimization as interactive intent inference over interpretable feature spaces under an information-theoretic framework, integrating principled query generation with visual interaction design for text-to-image generation.
- We demonstrate through systematic evaluations that APE achieves superior alignment with improved efficiency across both controlled benchmarks and user-defined tasks, showing that visual query interactions can serve as an effective and accessible complement to text-based prompting.

Our work demonstrates that reframing prompting as bi-directional communication, where systems actively elicit preferences rather than passively interpret text, can address alignment challenges that model improvements alone fall short of. By combining information-theoretic query selection with transparent, editable representations, APE makes text-to-image generation more accessible to general users. Source code is available at https://github.com/e-wxy/Adaptive-Prompt-Elicitation.

2 Related Work

Our work draws upon research across three primary areas: prompt optimization for text-to-image generation, interactive techniques in text-to-image systems, and preference learning from human feedback.

2.1 Prompt Engineering for Text-to-Image Generation

Early research established foundational principles for prompt design. Liu et al. [51] provided comprehensive guidelines on prompt keywords and model hyperparameters, while Oppenlaender [59] systematically categorized prompt modifiers through an ethnographic study, identifying six key types: subject terms, style modifiers, image prompts, quality boosters, repeating terms, and magic terms. These guidelines consistently suggest that more detailed and specific prompts tend to produce higher-quality images.

Complementing this line of work, automatic prompt optimization techniques have been proposed to help users write effective prompts without extensive learning. A prominent approach involves training specialized language models as prompt optimizers. Promptist [29] and BeautifulPrompt [7] combine supervised learning on curated preference data with reinforcement learning from AI feedback to improve semantic alignment and aesthetic quality. Prompt Expansion [15] and UF-FGTG [32] train deep generative models to transform input prompts into high-quality, detailed descriptions using large-scale prompt expansion datasets, thereby enhancing the aesthetic quality and diversity of generated images. More recently, large language models have been explored as blackbox prompt optimizers [10, 31, 50, 56], leveraging their natural language understanding capabilities to refine user inputs.

While manual prompt design offers more precise control and user agency, it suffers from steep learning curves and strong model-specific knowledge requirements, limiting scalability across users and use cases [28, 54]. Automatic prompt optimization reduces the need for user expertise but often overlooks the fundamental challenge of enabling users to express their intentions precisely, as well as the personalized and nuanced nature of user preferences [27, 54, 82].

2.2 Interaction Techniques in Text-to-Image Generation

Interaction techniques have emerged as a critical middle ground between manual prompt engineering and fully automatic optimization. By integrating human feedback into the generation loop, these techniques seek to bridge the gap between abstract user intent and concrete model outputs. We categorize prior work into three primary paradigms: interface-driven exploration, mixed-initiative control, and query-based ambiguity resolution.

To mitigate the trial-and-error nature of prompting, several systems provide structured interfaces for exploring the prompt space. Promptify [4], PromptMagician [21], and PrompTHis [25] support iterative refinement through prompt suggestions and history visualization. PromptMap [1] visualizes large collections of prompts as semantic maps, helping users navigate high-dimensional design spaces by clustering similar aesthetic outcomes. ThematicPlane [44] enables users to navigate and manipulate high-level semantic concepts within an interactive thematic design plane. While these tools reduce the memory burden of tracking iterations, they still rely on users to manually evaluate and synthesize multiple prompt candidates, which can become cognitively overwhelming as the search space grows.

Moving beyond simple text entry, recent work has introduced mixed-initiative systems that combine automated optimization with fine-grained user control. PromptCharm [84] exemplifies this approach by integrating an automated prompt optimizer (Promptist [29]) with explainable AI features, such as visualizing token-level attention to help users understand how specific words influence the generated image and enable targeted refinement. Other systems expand the interaction modalities: PromptPaint [13] and SketchFlex [48] allow users to constrain generation through spatial sketches, while AdaptiveSliders [36] enables continuous adjustment of semantic concepts. StyleFactory [92] facilitates style alignment by offering an interactive interface for granular style-strength adjustment and systematic evaluation. Brickify [73] supports expressive design intent specification through direct manipulation of design tokens. Although these methods offer expressive control, they often raise the barrier to entry, as interpreting attention maps or specifying spatial constraints may require technical expertise or artistic skills that novice users may lack.

Most closely related to our work, Proactive Agent [27] parses user prompts into belief graphs to identify underspecified attributes or entities, and generates clarifying questions to fill these gaps. Twin-Co [82] uses image captioning to detect discrepancies between user descriptions and generated outputs, then asks targeted questions to resolve mismatches. These pioneering works establish query-based interaction as a powerful paradigm for eliciting intent. However, query efficiency and the articulation burden of textual interaction remain underexplored. APE addresses these limitations by grounding query generation in information-theoretic principles and shifting from textual to visual interaction, enabling more efficient and accessible preference elicitation with potential to generalize to more complex professional design tasks.

2.3 Preference Learning from Human Feedback

Preference learning from human feedback has emerged as a powerful paradigm for aligning generative models with user intent. Reinforcement Learning from Human Feedback (RLHF) enables the training of models using comparative human feedback without requiring task-specific reward engineering [11]. Optimization methods such as Proximal Policy Optimization (PPO) [61, 71] and Direct Preference Optimization (DPO) [63] have been proposed to align language models with human preferences. Recent extensions of RLHF have been applied to image generation, where large-scale preference signals are leveraged to train text-to-image models, leading to improvements in aesthetic quality, prompt alignment, and visual fidelity [3, 45, 88].

Beyond offline population-level training, Interactive Machine Learning (IML) [18] facilitates real-time adaptation to user feedback. Given the high cost of human attention, sample efficiency is paramount. Information-theoretic frameworks, such as Bayesian experimental design, provide principled foundations for making optimal decisions in this context [9, 35, 64, 75]. Such principles underpin active learning strategies for acquiring labeled data [23, 33, 37, 39, 53], and manifest specifically as active preference learning, which focuses on efficiently eliciting user preferences through strategic query selection. Related ideas have been widely explored within Preferential Bayesian Optimization (PBO) [24], which aims to find

the optimum of black-box functions based on preference feedback, and applied to assist users in various design tasks, including visualization, photography, and typography [5, 41, 42, 78, 89]. BayesGenie [6] further integrates Bayesian Optimization with large language models (LLMs) to optimize image editing parameters based on user feedback. However, these methods typically operate over predefined continuous parameter spaces, limiting their applicability to open-ended tasks where user requirements are heterogeneous, not known a priori, and difficult to parametrize numerically.

More recently, LLM-based preference elicitation methods have emerged that use language models to learn user preferences through natural language dialogue [40, 46, 55], enabling more flexible design spaces and a more natural alignment with human intent.

Building upon these insights, this work extends active preference learning to text-to-image generation by grounding information-theoretic query selection in interpretable visual feature spaces. Unlike prior methods that operate over fixed low-dimensional parameter spaces [6], APE uses dynamically generated visual queries to reduce articulation burden while maintaining principled optimization of intent alignment.

3 Adaptive Prompt Elicitation

This section presents Adaptive Prompt Elicitation (APE), our interactive method for aligning text-to-image generation with user intent through strategic visual queries. We first illustrate the system through a walkthrough scenario, then formalize the alignment problem and describe our technical approach.

3.1 System Walkthrough

We introduce APE through a realistic scenario where Sarah, a hiking enthusiast with no prompt engineering experience, wants to create an app icon for her club's upcoming events. Sarah begins with a simple, underspecified prompt: *"an app icon for hiking."* Based on this initial input, the system presents a visual query — "What should be the main visual element?" — accompanied by several image options, including mountains, backpacks, boots, and trails (Figure 2A). Sarah selects the mountain option, which best aligns with her design intent. APE then updates its interpretation of her intent and generates the next visual query accordingly. If none of the options are satisfactory, Sarah can specify her preferences using the "Other" option.

Each selection incrementally refines Sarah's design intent, which is continuously reflected and updated in the requirements panel (Figure 2B). In addition to responding to system-generated visual queries, Sarah can directly add or modify the textual requirements in this panel to further articulate her desired outcome. Once the requirements are finalized, she clicks "Submit" (Figure 2C) to generate an image. This process can be repeated iteratively until she achieves a satisfactory result.

3.2 Problem Formulation

Let π represent a text-to-image model that maps prompts x to distributions over images y. Each user has a latent visual intent θ^* that characterizes their preferences across visual features like style, composition, mood, and subject matter. The goal is to find a prompt x^* that generates images maximally aligned with the user's true intent:

$$x^* = \arg\max_{x \in \mathcal{X}} \mathbb{E}_{y \sim \pi(\cdot|x)} [U_{\theta^*}(y)] \qquad (1)$$

where U_{θ^*} is the alignment utility function parameterized by θ^*. This formulation assumes access to the user's true intent θ^*, which is unknown in practice.

In manual prompt engineering, users attempt to learn a mapping $f_\pi : \Theta \rightarrow \mathcal{X}$ that translates latent intent θ^* into effective prompts, despite limited knowledge of the model π and the difficulty of expressing abstract visual concepts in natural language. This results in inefficient trial-and-error search in a high-dimensional prompt space.

Automatic prompt optimization instead learns a transformation $g_\pi : \mathcal{X} \rightarrow \mathcal{X}$ that refines an initial prompt. While effective at capturing population-level preferences through curated data, these methods may fail to adapt to individual users and assume that the initial prompt sufficiently reflects the user's latent intent.

Both paradigms suffer from complementary forms of information asymmetry: users do not know how the model interprets prompts, while the system does not know the user's true intent. We address these limitations via a bi-directional interactive optimization, which jointly optimizes prompts while adaptively eliciting θ^* through user interaction, bridging the gap between what users envision and what models produce.

3.3 System Overview

APE consists of three main components that work in concert to elicit user intent and generate optimized prompts (Figure 3): (1) **User Intent Modeling**, which represents and updates beliefs about user preferences; (2) **Adaptive Query Generation**, which selects informative visual queries based on current uncertainty; and (3) **Prompt Synthesis**, which translates elicited intent into effective prompts. We describe each component in detail below.

3.4 User Intent Modeling

The intent modeling component maintains and refines APE's understanding of the user's latent visual intent θ^* as responses are collected. We parameterize the user's intent through visual features. Let $\mathcal{V}$ denote the space of relevant visual features (e.g., "artistic style," "lighting," "color palette"), where each feature $v \in \mathcal{V}$ has an associated value domain (e.g., "impressionist" or "photorealistic" for style). The user's true intent θ^* specifies desired values for each feature. Rather than committing to a single hypothesis in the full feature space, APE represents uncertainty through a partial specification R_t and models plausible completions using language model priors.

Partial Intent Specification. At iteration t, APE maintains a structured specification $R_t = \{(v_i, s_i)\}$ consisting of feature-value pairs. Each pair (v_i, s_i) indicates that visual feature v_i (e.g., "artistic style", "lighting") should take value s_i (e.g., "watercolor", "golden hour"). Initially, R_0 summarizes features explicitly stated in the user input. As users respond to queries, R_t grows incrementally: $R_{t+1} = R_t \cup \{(v_{q_t}, a_t)\}$ where v_{q_t} is the queried feature and a_t is the user's selected value. This representation serves as a transparent interface between user and model, allowing the system to update its belief
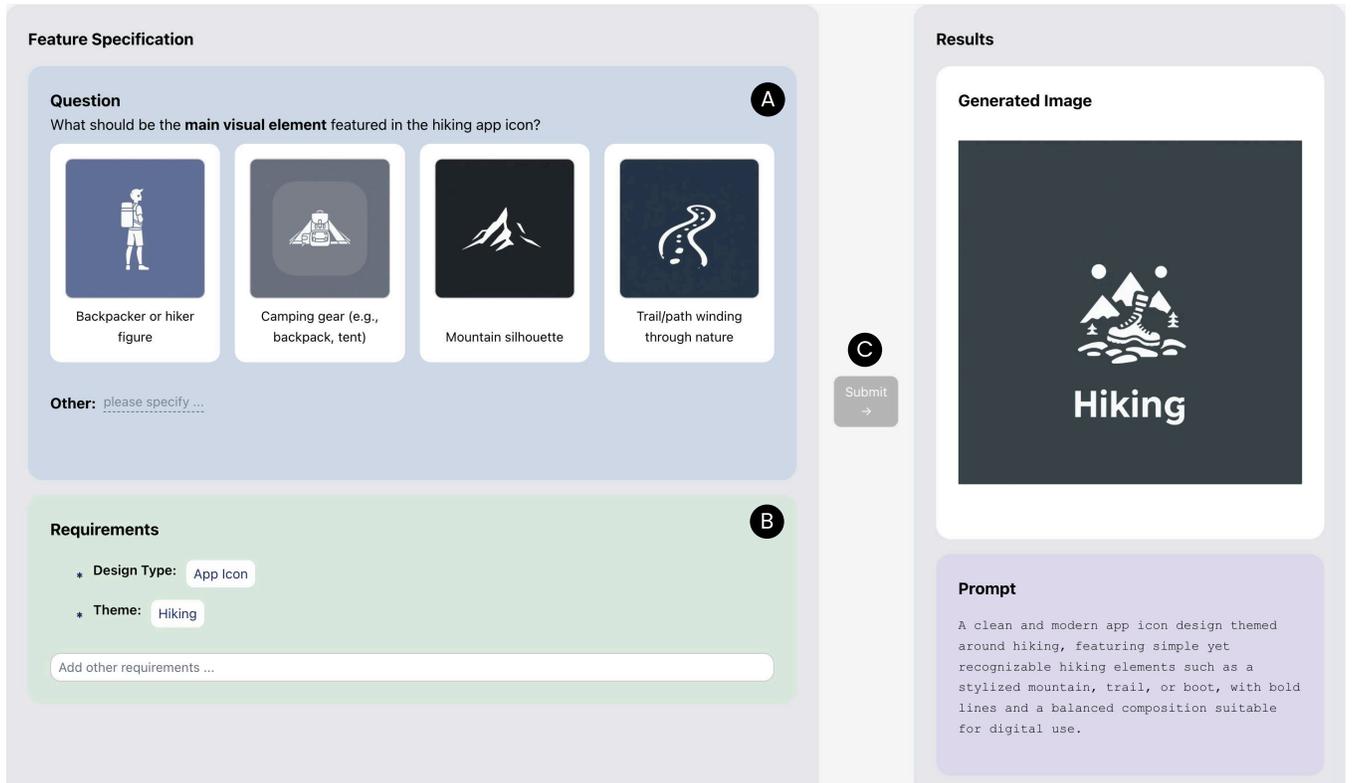
Figure 2: User interface design. The left area functions as feature specification, consisting of (A) question panel that presents visual queries, and (B) requirements panel that presents visual feature requirements specified during the interactions. (C) The user can generate an image by clicking the "Submit" button, which translates the requirements into effective prompts to generate an image.

adaptively. Additionally, it provides an interpretable trace of elicited preferences that users can review and modify if needed.

Dynamic Feature Space. To better adapt to the user's needs, instead of fixing the feature taxonomy a priori, APE's feature space $\mathcal{V}_t$ evolves dynamically based on context. Specifically, we utilize the LLM's in-context learning ability to propose relevant features conditioned on the current specification R_t. For example, if R_t contains (subject, "portrait"), the LLM might propose features like "facial expression", "gaze direction", or "background style" that are semantically relevant to portraits but not to landscape scenes. This adaptive approach ensures queries remain contextually appropriate and generalizable to various tasks. We prompt the LLM with R_t and request: "Given these specified visual features, what other important features would help clarify the user's vision?" The LLM returns a list of candidate features $\mathcal{V}_t = \{v_1, v_2, \ldots, v_n\}$, which forms the basis for query generation.

LLM Persona for Uncertainty Modeling. When R_t is underspecified, many complete intents θ remain consistent with the specified features. APE reasons about this uncertainty by modeling the distribution over plausible complete intents through the *LLM persona*, denoted $p_{\mathcal{L}}(\theta \mid R_t)$. The LLM persona captures prior knowledge

of visual intent based on the LLM's training on diverse visual descriptions.

Handling User Response. When the user selects value a_t for queried feature v_{q_t}, we update the specification to $R_{t+1} = R_t \cup \{(v_{q_t}, a_t)\}$ and update the posterior of user intent via $p_{\mathcal{L}}(\cdot \mid R_{t+1})$. This resampling incorporates the user's feedback, causing the LLM to generate samples consistent with both previous and newly specified features. Through iterative refinement, the sampled intents progressively converge toward the user's true latent intent θ^*.

3.5 Adaptive Query Generation

The query generation component determines which visual feature to query next. Although a straightforward approach that enumerates all features would yield complete knowledge of user intent, it risks expending user effort on irrelevant or low-impact dimensions. We formulate query generation as a decision-making problem within an information-theoretic framework [9, 35, 64], with the objective of proposing queries that are expected to maximally increase alignment utility.

Candidate Query Generation. At iteration t, we first generate a set of candidate queries Q_t from the dynamic feature space $\mathcal{V}_t$. For each candidate feature $v \in \mathcal{V}_t$ that has not yet been specified in
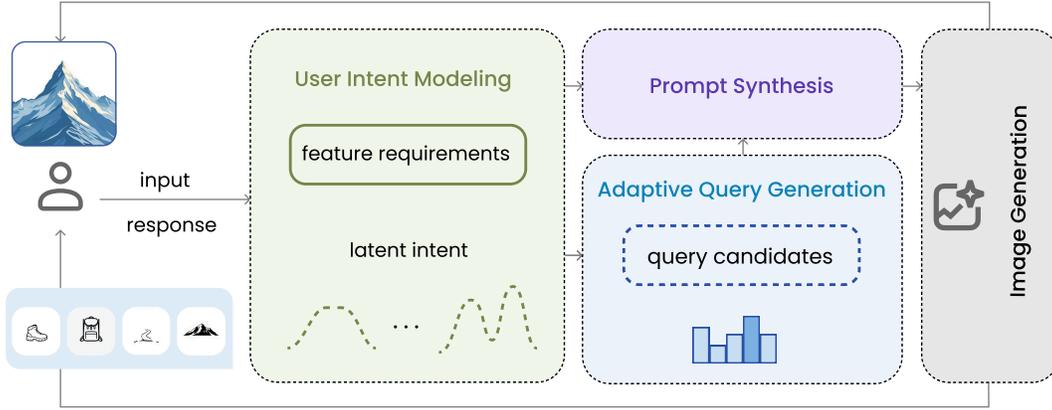
Figure 3: APE's computational pipeline. The system maintains a user intent model over visual features, adaptively generates visual queries to infer the user's intent, and then synthesizes optimized prompts for text-to-image generation models.

R_t, we generate diverse value options that partition the feature's domain. To ensure complete coverage when finite options cannot exhaustively represent all possibilities, we include a residual option ("Other") that allows users to indicate their preference lies outside the presented choices. If selected, this triggers a follow-up free-form query where users can specify their desired value in natural language.

Adaptive Query Selection. For each candidate query q about feature v_q with options $\{s_1, \ldots, s_m\}$, we define the Expected Alignment Utility Gain (EAUG) as:

$$\begin{aligned} \text{EAUG}(q; R_t) &= \alpha_{v_q} \cdot H[p(\cdot \mid v_q, R_t)] \\ &= -\alpha_{v_q} \sum_{i=1}^{m} p(s_i \mid v_q, R_t) \log p(s_i \mid v_q, R_t) \end{aligned} \quad (2)$$

where $H[p(\cdot \mid v_q, R_t)]$ denotes the Shannon entropy quantifying our uncertainty about which value the user prefers for feature v_q, which collapses to zero upon observing a definitive response, assuming user feedback is sufficiently reliable and can be calibrated. The weighting factor α_{v_q} captures the extent to which feature v_q influences overall alignment. For computational tractability, we approximate these quantities using the LLM persona samples $\{\theta^{(k)}\}_{k=1}^{K}$ drawn from the user intent model.

APE selects the query that maximizes the improvement in alignment quality:

$$q_t^* = \arg\max_{q \in Q_t} \text{EAUG}(q; R_t) \quad (3)$$

This strategy prioritizes queries about features that both have uncertain outcomes and substantially impact final alignment quality.

Visual Query Presentation. Once query $q_t^* = (v_{q_t}, \{s_1, \ldots, s_m\})$ is selected, APE generates visual exemplars to present each option. For each value option s_i, we:

(1) Construct a prompt that emphasizes the target feature-value pair: $x_i = f(\{(v_{q_t}, s_i)\} \mid R_t)$, where f is a prompt synthesis function as described in Section 3.6

(2) Generate an exemplar image: $y_i \sim \pi(\cdot \mid x_i)$ using the base text-to-image generation model

(3) Present images $\{y_1, \ldots, y_m\}$ side-by-side with descriptive labels

To emphasize the queried feature, we use a fixed random seed and identical generation parameters across all options, varying only the feature-specific prompt component. This controlled setup isolates the effect of the target feature while keeping other visual aspects consistent.

The query interface presents a simple multiple-choice question: "Which [feature name] best matches your vision?" followed by the labeled exemplar images. Users select their preferred option through a single click, instead of articulating preferences textually. This leverages recognition rather than recall, substantially reducing the articulation burden.

3.6 Prompt Synthesis

The prompt synthesis component translates the elicited intent specification R_T into an optimized natural language prompt. We formalize synthesis as a function $f : \mathcal{R} \to \mathcal{X}$ that maps explicit intents to text prompts.

$$x^* = f(R_T) = \mathcal{L}(m, c_\pi, R_T) \quad (4)$$

where $\mathcal{L}$ is a large language model, m is a meta-prompt that guides synthesis, and c_π is the context regarding the targeted image generation model. The meta-prompt m orchestrates synthesis in two stages: We first ask $\mathcal{L}$ to generate prompt engineering guidelines specific to the target model π. These elicited guidelines g encode model-specific best practices. Using g as context, we instruct $\mathcal{L}$ to synthesize a prompt incorporating all features from R_T.

This adaptive design ensures synthesis is both faithful to user intent and optimized for model-specific conventions. By explicitly eliciting guidelines, we make the LLM's prompt engineering knowledge transparent and consistent.

We now validate its effectiveness through systematic evaluation. The following two sections assess APE's performance through complementary lenses: controlled technical evaluation with ground-truth benchmarks (Section 4) establishes whether the approach achieves its core objectives of improving alignment and efficiency, while subsequent user studies (Section 5) examine how these technical capabilities translate to real-world user experience.

4 Technical Evaluation

To establish APE's technical capabilities, we conducted a controlled evaluation on benchmarked tasks with ground-truth references. This evaluation addresses fundamental questions about APE's core mechanisms: whether adaptive query elicitation improves alignment quality, and how efficiently information-theoretic query selection converges compared to alternative strategies. We employed LLM-based user simulators that responded to visual queries according to predefined targets, enabling us to (1) evaluate at scale across diverse tasks and (2) systematically compare strategies under identical conditions.

Our evaluation addresses two research questions:

- **RQ1**: Does APE improve alignment between user intent and generated images compared to existing prompt optimization approaches?
- **RQ2**: How efficiently does APE converge to high-quality prompts, and how does information-theoretic query selection compare to alternative strategies?

We evaluate these questions across two complementary benchmarks: DesignBench for users with visual preferences but limited articulation ability, and IDEA-Bench for users with clear conceptual goals requiring translation into effective prompts. Together, these benchmarks provide comprehensive coverage of text-to-image use cases and establish APE's technical foundations before examining real-world user experience in Section 5.

4.1 Method

4.1.1 Datasets. We evaluated APE on two complementary benchmarks that represent different user scenarios and task complexities:

- **DesignBench** [27]: comprises 30 diverse cases for daily usage, including hand-drawn cartoons, surreal photorealistic scenes, and artistic photographs. Each case provides an initial prompt, a target image, and a detailed description.
- **IDEA-Bench** [47]: encompasses comprehensive professional design and art tasks for image generation, distilled from real-world design cases across various platforms. From the complete dataset containing both text-to-image and image-to-image tasks, we selected 29 text-to-image generation cases spanning architectural style, business cards, sculptures, tickets, landmarks, logos, posters, and interior design. Each case includes a concise task description and a ground truth text prompt. We generated reference images using the ground truth prompts with FLUX.1 [43] (5 images per case to establish robust references).

Together, these benchmarks cover the spectrum from intuitive visual preferences to structured design requirements, providing comprehensive coverage of text-to-image use cases.

4.1.2 User Simulation. To enable systematic evaluation at scale while maintaining experimental control, we developed LLM-based user simulators that represent different user profiles:

- **Vision-Based Simulator (DesignBench).** We employed an LLM as a multi-modal user simulator that receives the target image as its "unarticulated vision." The simulator responds to visual queries by analyzing the reference image, mimicking users who know what they want when they see it but cannot describe it initially. This approach simulates the common scenario where users specify their preferences through visual feedback.
- **Intent-Based Simulator (IDEA-Bench).** For professional design tasks, we used an LLM simulator that adopts the ground truth prompt as its internal "creative brief." This configuration represents users with well-defined conceptual goals who need assistance translating intent into effective model input.

To ensure response quality and prevent hallucination, we designed simulation prompts using Chain-of-Thought reasoning [86], requiring simulators to analyze options before selection. To prevent leaking the ground truth to APE, only the question answer updates the feature specification R_t, with explanatory reasoning discarded.

While simulated users cannot fully replicate human preference formation or capture the complexity of human perception, they provide essential advantages for technical evaluation: perfect experimental control, reproducibility across conditions, elimination of learning effects, and the ability to test at scale across diverse scenarios. These controlled conditions allow us to isolate APE's technical contributions.

4.1.3 Baselines. We compare APE against three conditions:

- **Unoptimized**: The initial prompts, establishing the starting point and demonstrating overall improvement.
- **Automatic Prompt Optimization (APO)**: A non-interactive automatic optimization approach that uses language models to refine initial prompts provided by the user [29]. This isolates the value of interactive elicitation.
- **In-Context Query**: An interactive baseline that generates queries based on LLM in-context reasoning without principled information-theoretic selection. This isolates the contribution of our adaptive query selection strategy.

4.1.4 Evaluation Metrics. We employed multi-modal alignment metrics to comprehensively assess different aspects of generation quality:

Image-Image Similarity measures visual fidelity between generated and reference images using DreamSim [22], a perceptual similarity metric trained on human judgments, and DINOv2 [60], a self-supervised vision transformer capturing semantic visual features. These metrics assess whether the generated images visually resemble the intended target.

Text-Text Similarity evaluates prompt quality by measuring semantic similarity between synthesized prompts and ground truth descriptions. We used E5 [83], a state-of-the-art text embedding model with strong cross-task transfer, and OpenAI's text-embedding-3-large [58]. High text-text similarity indicates that APE successfully elicits and encodes the intended semantic content.

Text-Image Similarity assesses cross-modal consistency between generated images and textual descriptions using VQAScore [49], which leverages visual question answering models to compute alignment by measuring the probability of affirmative responses to "Does this figure show [text]?" This metric captures whether the visual output matches the linguistic specification.

This multi-metric approach captures complementary aspects of alignment: visual fidelity (image-image), semantic correctness (text-text), and cross-modal consistency (text-image).

4.1.5 Apparatus. Our evaluation employed the following configuration: We used the FLUX.1-schnell text-to-image model [43] accessed via the Fal API[1] with eight inference steps, providing a balance between generation quality and computational efficiency. The user simulator is based on GPT-4o-mini [58], accessed through the OpenAI API. Each simulation proceeded for a maximum of 15 iterations, ensuring sufficient exploration while maintaining realistic interaction limits. In each round, up to five candidate queries were generated, each with a maximum of five options, offering diverse yet cognitively manageable choices. Finally, we performed five independent experimental runs per case to establish reliable confidence intervals.

4.2 Improved Alignment Quality (RQ1)

4.2.1 Overall Performance. Table 1 presents comprehensive alignment results across both datasets. APE achieves the best performance in 8 of 10 metric-dataset combinations, demonstrating consistent improvements across diverse scenarios and evaluation modalities. This multi-metric consistency suggests that APE effectively bridges the gap between user intent and model interpretation.

For IDEA-Bench's challenging professional design scenarios, APE demonstrates substantial gains over the baseline. Image similarity metrics show marked improvements (DINOv2: +30.2%, DreamSim: +12.1%), indicating that interactive elicitation provides considerable value for complex creative tasks requiring domain-specific vocabulary and nuanced specifications. These gains are particularly pronounced compared to the more modest improvements observed on DesignBench, suggesting that APE's benefits scale with task complexity.

Comparing APE and In-Context Query against APO reveals the critical value of bidirectional communication. APO shows minimal improvement over baseline across both datasets, highlighting a fundamental challenge of non-interactive approaches: discovering user preferences that remain unexpressed in the initial prompt. Even sophisticated language models operating on user input alone lack sufficient signal to infer latent visual intentions, underscoring the necessity of explicit preference elicitation.

APE's advantages over In-Context Query demonstrate the practical benefits of principled query generation. While both methods engage users interactively, APE's information-theoretic approach systematically identifies more informative and impactful features, translating to measurably better alignment outcomes. This advantage appears across multiple metrics and task types, validating the theoretical foundation of our query selection strategy.

4.2.2 Task-Specific Performance. Table 2 reveals heterogeneous performance across IDEA-Bench's nine design categories, providing insights into how task characteristics interact with our elicitation approach.

Three categories demonstrate substantial gains: Interior Design (DINOv2: +85.4%, DreamSim: +22.7% over baseline), Poster Design (DINOv2: +83.7%, DreamSim: +8.6%), and Logo Design (DINOv2: +59.3%, DreamSim: +18.3%). These tasks require coordinating multiple interdependent visual elements where initial prompts underspecify numerous critical attributes. Architectural Style (DINOv2: +29.6%, DreamSim: +2.0%) and Painting (DINOv2: +18.1%, DreamSim: +8.3%) show moderate gains with notable metric divergence—substantial DINOv2 improvements coupled with minimal DreamSim changes may reflect limitations in DreamSim's ability to detect fine-grained visual refinements. Landmark Building shows modest gains (DINOv2: +7.2%, DreamSim: +3.4%), with APE and In-Context Query performing comparably. We hypothesize this occurs because models possess strong priors for well-documented landmarks from training data, making additional elicitation less critical. Interestingly, In-Context Query (DINOv2: 0.577) slightly outperforms APE (DINOv2: 0.499) in Business Card design. This anomaly suggests that for highly constrained tasks with limited degrees of freedom, strategic query selection may provide minimal advantage over simpler interactive approaches.

It is important to note that image similarity metrics may not fully capture fine-grained professional design quality. For Sculpture, where the underlying text-to-image model is known to struggle with geometrically accurate three-dimensional forms, and Ticket Design, where the model notoriously fails at rendering coherent text [47], similarity metrics may primarily capture whether the correct object category is present rather than professional design quality such as anatomical accuracy or typography legibility. The observed improvements in these categories, while statistically meaningful, should be interpreted with caution as metric gains may not correspond to perceptually meaningful quality differences for humans.

4.3 Improved Query Efficiency (RQ2)

Figure 4 tracks alignment quality evolution across interaction rounds, revealing APE's superior convergence properties. On both datasets, APE reaches higher final alignment scores while requiring fewer interactions, with efficiency gains particularly pronounced in early iterations. This pattern suggests that APE's principled query strategy identifies high-impact features during this critical early phase, enabling rapid convergence toward user intent. This efficiency advantage is more pronounced on IDEA-Bench, suggesting that professional design tasks with richer feature spaces amplify the value of strategic query selection. APE demonstrates steeper initial improvement followed by gradual refinement, indicating that it effectively prioritizes foundational design decisions—those with broad downstream impact—before addressing fine-grained details.

The observed efficiency gains translate to meaningful practical benefits. Reducing the number of required interactions decreases both user effort and cumulative interaction time. For text-to-image systems where each generation cycle incurs computational cost and

---

Table 1: Alignment quality on DesignBench and IDEA-Bench (mean ± 95% CI). Best results per row bolded.

Dataset	Modality	Metric↑	Unoptimized	APO	In-Context	APE (Ours)
DesignBench	Text-Text	E5	$0.887_{\pm 0.05}$	$0.894_{\pm 0.05}$	$0.898_{\pm 0.02}$	$\mathbf{0.899}_{\pm 0.02}$
		OpenAI	$0.732_{\pm 0.12}$	$0.789_{\pm 0.10}$	$\mathbf{0.811}_{\pm 0.04}$	$0.797_{\pm 0.04}$
	Image-Image	DINOv2	$0.664_{\pm 0.32}$	$0.668_{\pm 0.33}$	$0.676_{\pm 0.15}$	$\mathbf{0.679}_{\pm 0.16}$
		DreamSim	$0.639_{\pm 0.16}$	$0.633_{\pm 0.15}$	$0.636_{\pm 0.08}$	$\mathbf{0.654}_{\pm 0.08}$
	Text-Image	VQAScore	$0.891_{\pm 0.14}$	$0.890_{\pm 0.15}$	$0.901_{\pm 0.04}$	$\mathbf{0.904}_{\pm 0.03}$
IDEA-Bench	Text-Text	E5	$0.888_{\pm 0.06}$	$0.886_{\pm 0.04}$	$\mathbf{0.903}_{\pm 0.01}$	$0.901_{\pm 0.02}$
		OpenAI	$0.785_{\pm 0.09}$	$0.758_{\pm 0.12}$	$0.788_{\pm 0.04}$	$\mathbf{0.792}_{\pm 0.04}$
	Image-Image	DINOv2	$0.471_{\pm 0.32}$	$0.513_{\pm 0.30}$	$0.596_{\pm 0.14}$	$\mathbf{0.613}_{\pm 0.14}$
		DreamSim	$0.554_{\pm 0.21}$	$0.557_{\pm 0.18}$	$0.611_{\pm 0.09}$	$\mathbf{0.621}_{\pm 0.09}$
	Text-Image	VQAScore	$0.675_{\pm 0.23}$	$0.675_{\pm 0.23}$	$0.674_{\pm 0.11}$	$\mathbf{0.678}_{\pm 0.10}$

Table 2: Task-specific alignment quality on IDEA-Bench across design categories (mean ± 95% CI). Best results per row bolded.

Design Type	Metric↑	Unoptimized	APO	In-Context	APE (Ours)
Architectural Style	DINOv2	$0.432_{\pm 0.07}$	$0.481_{\pm 0.08}$	$0.503_{\pm 0.10}$	$\mathbf{0.560}_{\pm 0.06}$
	DreamSim	$0.586_{\pm 0.04}$	$0.576_{\pm 0.01}$	$0.520_{\pm 0.03}$	$\mathbf{0.598}_{\pm 0.02}$
Business Card	DINOv2	$0.474_{\pm 0.09}$	$0.453_{\pm 0.07}$	$\mathbf{0.577}_{\pm 0.09}$	$0.499_{\pm 0.11}$
	DreamSim	$0.477_{\pm 0.05}$	$0.471_{\pm 0.04}$	$\mathbf{0.527}_{\pm 0.03}$	$0.508_{\pm 0.02}$
Interior Design	DINOv2	$0.362_{\pm 0.04}$	$0.619_{\pm 0.08}$	$0.622_{\pm 0.05}$	$\mathbf{0.671}_{\pm 0.07}$
	DreamSim	$0.576_{\pm 0.04}$	$0.634_{\pm 0.03}$	$0.684_{\pm 0.01}$	$\mathbf{0.707}_{\pm 0.02}$
Painting	DINOv2	$0.386_{\pm 0.06}$	$0.372_{\pm 0.06}$	$0.451_{\pm 0.03}$	$\mathbf{0.456}_{\pm 0.05}$
	DreamSim	$0.555_{\pm 0.06}$	$0.554_{\pm 0.07}$	$0.577_{\pm 0.05}$	$\mathbf{0.601}_{\pm 0.05}$
Sculpture	DINOv2	$0.601_{\pm 0.07}$	$0.611_{\pm 0.04}$	$0.667_{\pm 0.04}$	$\mathbf{0.737}_{\pm 0.02}$
	DreamSim	$0.544_{\pm 0.02}$	$0.545_{\pm 0.03}$	$\mathbf{0.635}_{\pm 0.02}$	$0.630_{\pm 0.04}$
Ticket Design	DINOv2	$0.560_{\pm 0.04}$	$\mathbf{0.561}_{\pm 0.06}$	$0.551_{\pm 0.05}$	$0.545_{\pm 0.06}$
	DreamSim	$0.570_{\pm 0.03}$	$0.580_{\pm 0.01}$	$0.609_{\pm 0.05}$	$\mathbf{0.621}_{\pm 0.03}$
Landmark Building	DINOv2	$0.788_{\pm 0.04}$	$0.813_{\pm 0.05}$	$\mathbf{0.854}_{\pm 0.01}$	$0.845_{\pm 0.02}$
	DreamSim	$0.756_{\pm 0.02}$	$\mathbf{0.793}_{\pm 0.02}$	$0.768_{\pm 0.01}$	$0.782_{\pm 0.03}$
Logo Design	DINOv2	$0.396_{\pm 0.03}$	$0.567_{\pm 0.05}$	$0.625_{\pm 0.03}$	$\mathbf{0.631}_{\pm 0.06}$
	DreamSim	$0.480_{\pm 0.03}$	$0.489_{\pm 0.03}$	$\mathbf{0.574}_{\pm 0.05}$	$0.568_{\pm 0.08}$
Poster Design	DINOv2	$0.319_{\pm 0.06}$	$0.465_{\pm 0.06}$	$0.548_{\pm 0.02}$	$\mathbf{0.586}_{\pm 0.03}$
	DreamSim	$0.545_{\pm 0.03}$	$0.569_{\pm 0.02}$	$\mathbf{0.605}_{\pm 0.02}$	$0.592_{\pm 0.03}$

user attention, this efficiency improvement represents substantial value in real-world deployment scenarios.

5 User Evaluation

While the technical evaluation (Section 4) demonstrates APE's performance on controlled benchmarks with ground-truth intents, real-world deployment requires understanding how the system performs with actual users pursuing open-ended creative goals. To address this, we conducted a between-subjects user study with 128 participants recruited from Prolific[2]. Unlike the technical evaluation, where ground truth was available for objective comparison, the user study focuses on subjective alignment quality, cognitive workload, and interaction efficiency as perceived by users with varying levels of prior experience engaging with self-defined creative tasks.

The study evaluates APE using a prototyped user interface (Figure 2) that reflects how the system would be deployed in practice, and compares it against manual prompt engineering, a baseline condition representing the current standard in which users iteratively refine textual prompts. This comparison is guided by three research questions:

- **RQ3:** How do users perceive the alignment of the generated image with their design intent when using APE compared to manual prompt engineering?
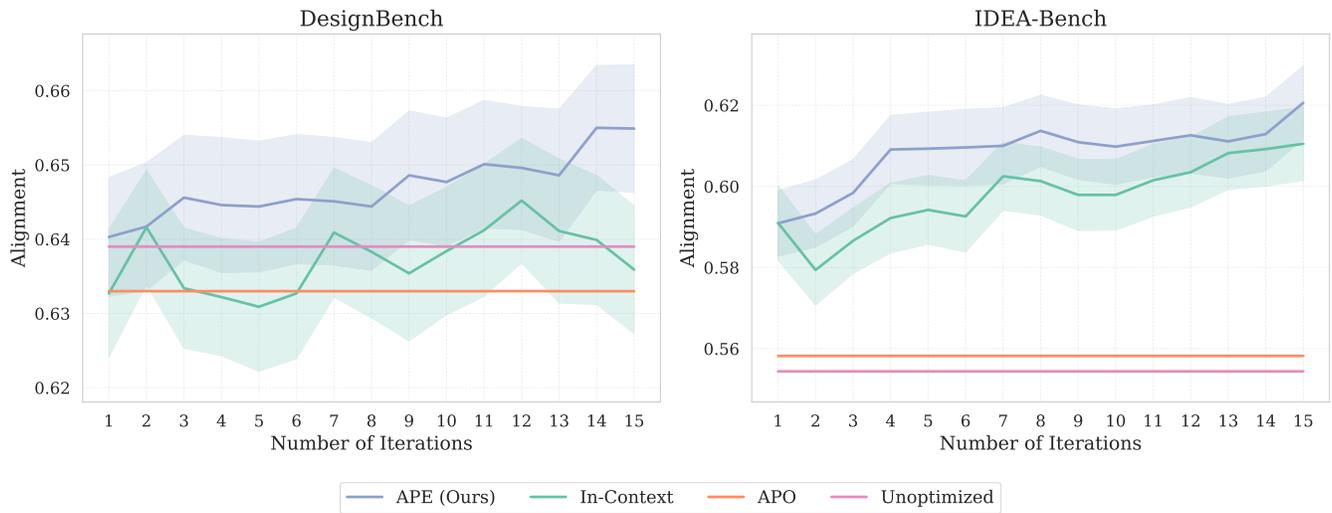
Figure 4: The progress of alignment over iterations, evaluated by image-image similarity (DreamSim↑) with 95% CI. APE demonstrates faster convergence and higher final alignment compared to In-Context Query.

- **RQ4:** How does APE affect users' perceived cognitive workload relative to traditional prompting approaches?
- **RQ5:** How does APE affect interaction efficiency in terms of iterations required to achieve aligned results?

These questions complement the technical evaluation by examining whether APE's computational advantages translate to meaningful improvements in user experience. The following subsections detail our methodology, present quantitative and qualitative findings.

5.1 Method

5.1.1 Participants. We recruited 128 participants from Prolific meeting the following criteria: at least a 99% historical approval rate, more than 100 historical submissions, proficiency in English, and having normal or corrected-to-normal vision. The sample included 48 female, 74 male, and 6 non-binary participants (M_{age} = 34.4, SD = 10.8). We did not filter by design expertise, aiming to capture a naturalistic distribution of users who might encounter these tools. This resulted in a diverse sample spanning varied levels of prior experience with LLMs, text-to-image systems, and design activities (see Table 4). Each session lasted approximately 25 minutes, and participants received £4 as compensation.

5.1.2 Task and Apparatus. Each participant was given a task scenario from a set of eight design types: interior design, painting, photography, app icon design, poster design, logo design, fashion design, and architectural style design. Given the scenario, participants decided their design idea and interacted with either our method or the baseline method to generate an image. The task descriptions are shown in Table 3. We selected these scenarios because they combine commonly used benchmarks in text-to-image generation research with familiar, everyday creative tasks. By covering a wide range of contexts—from artistic creation (painting, fashion design) to practical design (logos, app icons, posters) and

imaginative visualization (dream interiors, ideal architecture)—we ensured both task diversity and participant familiarity. This balance encourages natural engagement and enables us to assess how generative models perform across a broad spectrum of creative objectives.

The user study was deployed on a web-based interface. The server side integrated our APE algorithm, supporting real-time interaction with the text-to-image model.

5.1.3 Experimental Design. We compared two conditions:
- **Baseline (Manual Prompt Engineering).** Participants could freely edit prompts and iteratively generate images. They could refine the prompt as many times as they wished within a 15-minute limit. This setup mirrors commonly used text-to-image tools.
- **APE (ADAPTIVE PROMPT ELICITATION).** Participants completed the task using our proposed APE system. Instead of manually editing prompts, the system engaged participants through adaptive, informative questions and automatically translated their elicited requirements into prompts. The interaction time was matched to the Baseline condition (15 minutes).

To avoid learning and ordering effects, we adopted a between-subjects design where each participant was assigned one condition and one task. Each task was used equally often per condition. We conducted a power analysis using G*Power [20] for a two-condition between-subjects design. Assuming a moderate effect size (f = 0.5), α = 0.05, and statistical power of 0.8, the required sample size was estimated at 128 participants.

5.1.4 Procedure. The experiment followed a structured sequence:
(1) **Introduction.** After reading an introduction to the study, participants provided informed consent and completed a pre-task survey assessing their prior experience with LLM tools, text-to-image systems, and design-related activities.

They then received a short tutorial on the assigned system, followed by a brief comprehension quiz.

(2) **Generation Task.** Participants were assigned one of the eight generation tasks and were encouraged to think about their desired outcome. They then used the assigned system for up to 15 minutes to generate images, with the option to stop earlier once being satisfied with the result.

(3) **Post-task Questionnaires.** After completing the task, participants filled out questionnaires assessing the alignment, task workload and effectiveness of questions.

5.1.5 Measurements. We measured three aspects: **alignment**, **workload** and **efficiency**.

Alignment. Since no standardized measure exists for assessing AI-alignment with user intentions in open-ended tasks, we deployed a custom questionnaire that draws conceptually from prior discussion of alignment gaps in these scenarios [77]. The questionnaire focuses on four dimensions: the *capability gap*, the *instruction gap*, and the *intentionality gap*, along with an *overall* assessment of envisioning-support of the system. For each dimension, participants responded to one pair of positively and negatively valenced statements reflecting potential system support or limitations (see Appendix A.3.1). All items were rated on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree), and negatively-worded items were reverse-scored before averaging across each dimension.

Task workload. was measured using the NASA Task Load Index (NASA-TLX) [30], a widely validated multidimensional measure of subjective workload.

Efficiency. was measured through:

- **Iteration Count:** The number of iterations required for participants to obtain a satisfactory image.
- **Useful Question Ratio:** The proportion of visual questions generated by the system that participants rated as useful.

Open-ended feedback. In addition, we collected open-ended, qualitative feedback about participants' experiences.

5.1.6 Data Analysis. We employed a mixed-methods approach. For quantitative measures, we used non-parametric Mann–Whitney U tests [57] with Bonferroni correction for multiple comparisons. For qualitative feedback, we conducted a thematic analysis [34]. We report key themes supported by representative participant quotes.

5.2 Perceived Alignment Quality (RQ3)

5.2.1 Quantitative Results. Figure 5 presents participant ratings across four dimensions of alignment between their envisioned concepts and the generated images. APE demonstrated statistically significant improvements across all four dimensions (Mann-Whitney U tests, Bonferroni-corrected, all $p < 0.05$). For capability gap, participants rated APE significantly higher ($M = 5.48, SE = 0.15$) than Baseline ($M = 4.45, SE = 0.21; U = 1309.5, p = 0.0016, r = 0.36$), indicating that APE's interactive elicitation process helped participants feel more capable of achieving their desired outcomes. The instruction gap showed the strongest improvement, with APE ($M = 5.85, SE = 0.07$) substantially outperforming Baseline ($M = 5.09, SE = 0.11; U = 1033.0, p < 0.0001, r = 0.50$). This suggests
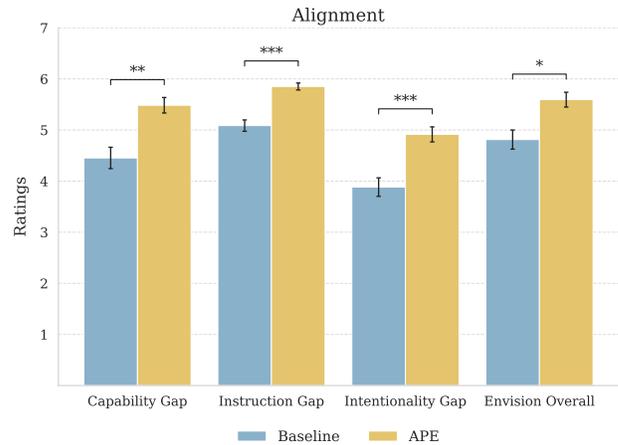

Figure 5: Alignment ratings (± SE) for four dimensions across conditions. Participants using APE reported significantly higher scores for all dimensions compared to Baseline. Asterisks indicate significance levels (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

that APE effectively helps users communicate preferences they struggle to articulate textually. The intentionality gap showed marked improvement with APE ($M = 4.91, SE = 0.15$) over Baseline ($M = 3.88, SE = 0.18; U = 1205.0, p = 0.0002, r = 0.41$). This gap of over one full rating point suggests that APE's structured elicitation process helps users clarify their own intent, enabling them to evaluate the result better. Finally, envision overall ratings confirmed this pattern, with APE ($M = 5.59, SE = 0.15$) significantly exceeding Baseline ($M = 4.81, SE = 0.19; U = 1428.5, p = 0.0118, r = 0.30$).

Computing overall alignment as the sum of ratings across all four dimensions, APE achieved significantly higher total scores ($M = 21.84, SE = 0.41$) compared to Baseline ($M = 18.23, SE = 0.58$), corresponding to a 19.8% improvement. These results provide convergent evidence that APE addresses a genuine need: helping users bridge the gap between their visual intentions and textual articulation. The consistent improvements across all alignment dimensions, with a medium effect size, suggest the approach successfully reduces both communication barriers and expectation mismatches.

5.2.2 Qualitative Feedback. Participants' primary complaints in the Baseline condition centered on the lack of alignment between their intent and the generated outputs, as well as the difficulty of achieving high-quality, fine-grained designs. Many reported that the system failed to fully interpret their prompts even after repeated refinements: "*I keep adding [to the] prompt, but the generated image is not getting better at understanding the prompt.*" Participants expressed frustration that the Baseline could not support detailed or precise adjustments: "*I was only about 65 percent satisfied because even though it made the image I was going for, it wouldn't let me tweak it further to take it to perfection.*" Additionally, users frequently observed inconsistent adherence to instructions—such as incorrect numbers of people, unstable elements across iterations,

or ignored constraints—which further undermined trust in the generation process: "*It felt like some of the prompts I wrote were ignored. Thus the image I had in mind seemed to be impossible to be created.*"

By contrast, in the APE condition, participants often described stronger alignment with their envisioned outcomes, crediting the elicited visual queries for surfacing overlooked or implicit design details: "*The questions presented to me were very helpful!*" and "*My first prompt wasn't the best, but with the option to make edits, my vision was brought to life.*" Several noted that the final images closely matched their mental image. However, a minority still reported occasional inconsistencies in low-level details (e.g., colors, small elements) or regressions during refinement, and some wished for greater output variety or finer-grained control after elicitation. This residual inconsistency, similar to what was observed in the Baseline condition, may stem from inherent limitations of the current text-to-image models rather than the elicitation approach itself.

5.2.3 The Effect of Prior Experience. Figure 6 presents alignment ratings across participants with varying levels of experience using text-to-image generation models. APE shows consistently high ratings across all experience levels, while Baseline exhibits a positive trend with increasing experience.

To examine whether prior experience modulated the effect of APE on alignment ratings, we fitted an ordinal logistic regression model with condition (Baseline vs. APE), prior experience (modeled as an ordinal variable based on self-reported usage frequency), and their interaction. The model was fitted using maximum likelihood estimation ($N = 128$, AIC = 917.9, BIC = 1023). The ordinal logistic regression revealed a significant main effect of condition ($\beta = 2.26$, $SE = 0.81$, $z = 2.80$, $p = 0.005$), with participants using APE reporting higher alignment ratings than those in the Baseline condition. Prior experience was positively associated with alignment ratings, although this effect did not reach statistical significance ($\beta = 0.32$, $SE = 0.18$, $z = 1.75$, $p = .080$). The interaction between condition and experience was not significant ($\beta = -0.29$, $SE = 0.27$, $z = -1.08$, $p = .280$), providing no evidence that the beneficial effect of APE on perceived alignment differed significantly across experience levels.

Complementarily, we conducted an exploratory analysis to further characterize how prior experience related to alignment ratings within each condition. We dichotomized participants based on self-reported usage frequency of text-to-image generation models into *low-experience users* (never used through monthly; $n = 93$) and *high-experience users* (weekly or more; $n = 35$). In the Baseline condition, high-experience users reported higher alignment ratings ($M = 19.73$, $SD = 3.78$) than low-experience users ($M = 17.56$, $SD = 4.87$), corresponding to a mean difference of 2.17 points. This difference approached conventional significance thresholds under a Welch-corrected t-test ($t(62) = 1.94$, $p = .059$) and was associated with a moderate effect size (Cohen's $d = 0.48$). However, given the limited observed power (0.41), this result should be interpreted cautiously as suggesting a potential association between prior experience and higher alignment ratings when using the Baseline system. In contrast, within the APE condition, alignment ratings were highly similar between high-experience users ($M = 21.80$, $SD = 3.52$) and low-experience users ($M = 21.86$, $SD = 3.24$), yielding a negligible mean difference of 0.06 points ($t(62) = 0.06$, $p = .956$, Cohen's $d = 0.02$). These exploratory results provide no

clear evidence that prior experience substantially modulated APE's effectiveness given our current sample size.

5.3 Cognitive Workload (RQ4)

5.3.1 Quantitative Results. Figure 7 presents NASA-TLX ratings across six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration. We did not find statistically significant differences between APE and Baseline across any dimension (Mann-Whitney U tests, Bonferroni-corrected, all $p > 0.05$).

These results suggest that APE achieves its alignment improvements without adding significant cognitive burdens. While one might hypothesize that answering multiple visual queries would increase mental demand or frustration compared to direct prompting, APE users reported statistically similar experiences across workload dimensions.

5.3.2 Qualitative Feedback. In the Baseline condition, users generally reported an acceptable cognitive load when expressing their intent in a single prompt. However, the major source of difficulty stemmed from the need to craft keyword-rich, model-aware prompts, which often led to iterative trial-and-error and substantial mental overhead. Representative comments included: "*I had to try many variants; long prompts made results worse*" and "*It was frustrating to want one thing and get a completely different result.*" (See Figure 14)

In the APE condition, many participants emphasized that APE reduced linguistic effort by presenting lightweight visual choices instead of requiring exhaustive textual descriptions: "*The visual prompts were extremely helpful*" and "*It helped me tease out details I had not thought of.*" Several participants explicitly reported lower workload and greater ease of use: "*It was very easy to use the system.*" Nonetheless, a few participants noted the desire for additional flexibility, such as the ability to skip or undo queries, or found some question sequences somewhat lengthy: "*I wish it was easier to go back though or to skip a question.*" A small fraction also reported feeling constrained when their intent could not be fully captured by the available visual options, requiring them to specify additional preferences manually.

5.4 Interaction Patterns and Efficiency (RQ5)

Beyond subjective ratings, we analyzed objective behavioral measures to understand how users actually interacted with APE and the resulting efficiency. We analyzed logged interaction data from all 64 participants in the APE condition, examining feature usage patterns, query effectiveness, and iteration requirements.

5.4.1 Interaction Patterns. Across all APE sessions, we logged 1,199 total user actions, distributed as follows: 645 visual query responses by selecting from provided options ("Select Answer Option" 53.8%), 31 visual query responses by providing "Other" answers ("Provide Other Answer", 2.6%), 367 prompt generation requests ("Generate Prompt", 30.6%), 98 requirement additions ("Add Requirement", 8.2%), and 58 requirement modifications ("Modify Requirement", 4.8%).

Figure 8 visualizes interaction flows across iterations. The distribution exhibits a front-loading effect, with approximately 68.89% of all interactions occurring within the first 15 iterations, after
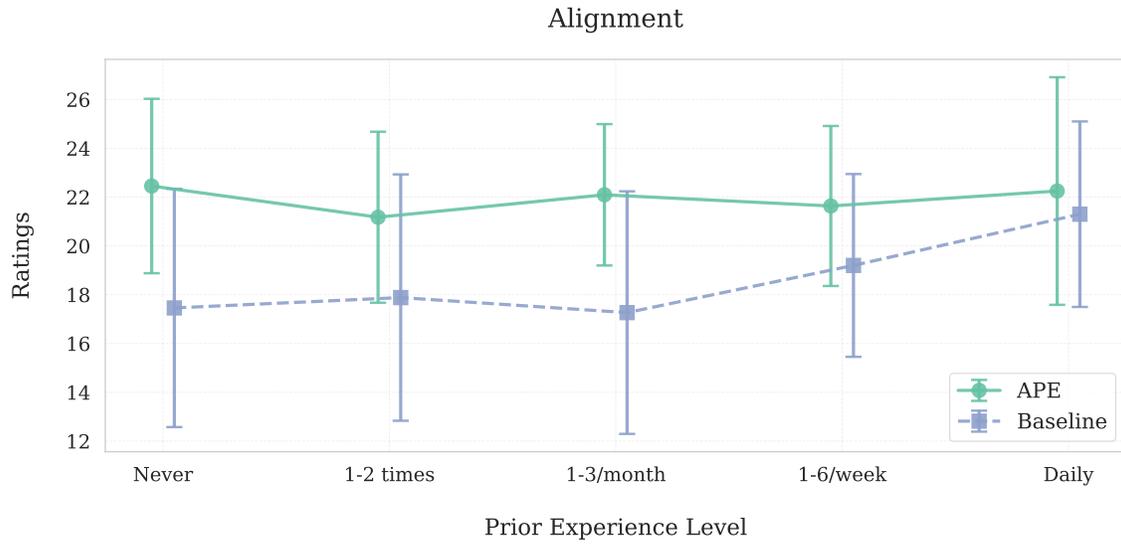
Figure 6: Alignment ratings (sum ± SD) by condition and prior experience level.
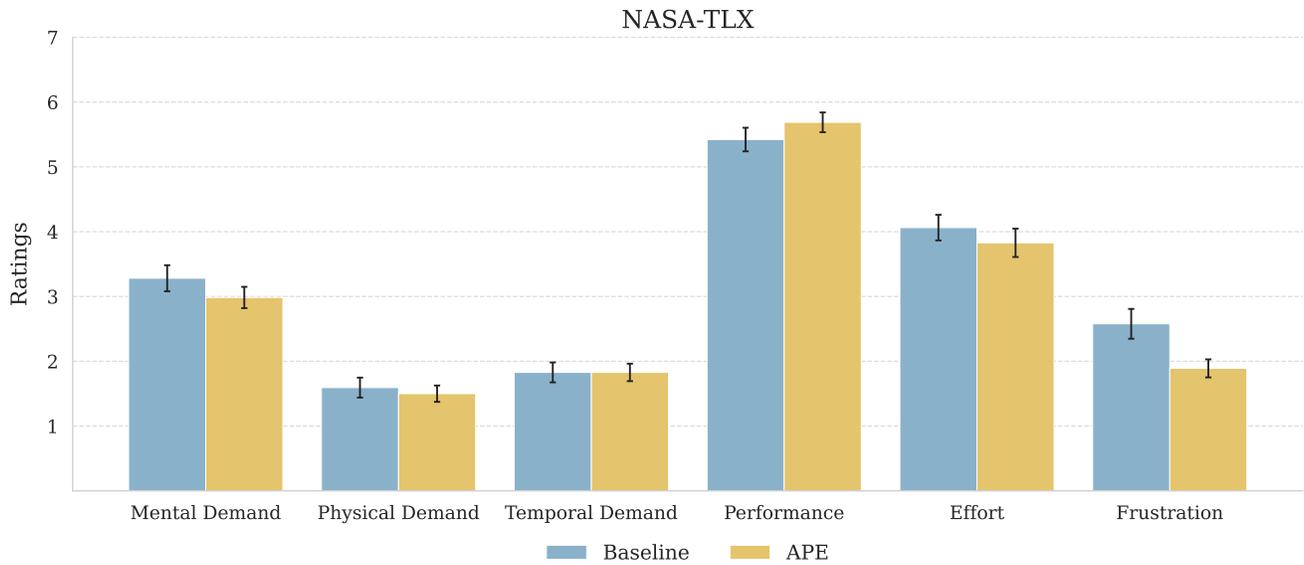

Figure 7: NASA-TLX ratings (± SE) across six dimensions for Baseline and APE conditions. No significant differences were found between conditions (all $p > 0.05$).

which interaction frequency steadily decreases. The figure shows a gradual drop-off in engagement over successive iterations.

Across iterations, users most commonly engage in repeated cycles of responding to visual outputs by selecting an answer option, followed by a transition to image generation. While answer-to-answer transitions dominate early iterations, the relative proportion of transitions leading to generation increases over time, indicating that users tend to move from exploratory visual queries to generation after several rounds of refinement. Detours into manual

requirement editing are comparatively infrequent. This pattern suggests that APE's visual query interactions help users iteratively refine intent before committing to generation, reducing reliance on repeated text-based prompt rewriting.

The rarity of certain interaction types provides additional insights. Participants selected the "Other" option in only 4.59% of visual query responses, with 46 participants never using this option and only 10 participants using it more than once. This low frequency suggests that LLM-generated feature options are capable
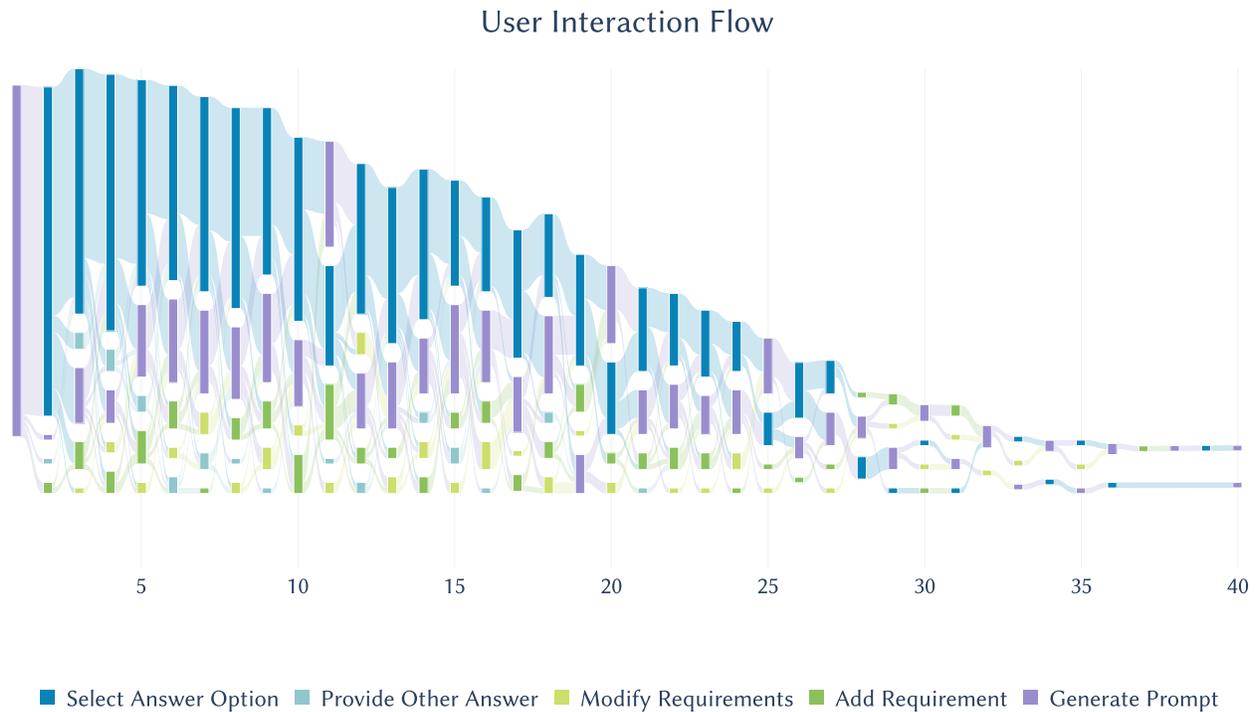
User Interaction Flow

■ Select Answer Option ■ Provide Other Answer ■ Modify Requirements ■ Add Requirement ■ Generate Prompt

Figure 8: Sankey diagram of user interaction flows across iterations (*n*=64 participants, 1,199 actions). Flow width indicates action frequency. Interactions are front-loaded, with 68.89% occurring within the first 15 iterations, followed by a steady decline in later iterations. Users commonly engage in several rounds of visual query responses before transitioning to image generation (purple). Transitions involving manual requirement editing are comparatively infrequent, suggesting that visual query interactions support iterative refinement prior to generation.

of capturing most users' preferences within the presented choices; however, it may also reflect a usability effect, where providing custom inputs requires greater effort and thus discourages use of the Other" option. Similarly, direct editing of the requirements panel occurred in only 13.01% of all interactions, with the vast majority of user effort directed toward responding to system-generated queries rather than manual text refinement.

The temporal concentration of activity also reveals differentiated user behaviors. While most (76.56%) users converged within 25 iterations, a small subset (9.38%) engaged beyond iteration 30, suggesting individual differences in task complexity and exploration preferences. Nevertheless, the overall pattern demonstrates that APE successfully guides users through a coherent, efficient workflow centered on visual query interaction rather than linguistic articulation.

Example of Interaction Trace. Figure 9 illustrates APE's progressive elicitation through a representative interior design task. The participant began with a concise specification: *"A large bedroom with open spaces, and that includes a fireplace, four-poster bed, big windows and lots of natural light."* While this 21-word prompt establishes the basic theme, it leaves numerous design decisions unspecified, such as aesthetic style, color palette, materials, and detailed furnishing choices.

Through six generation rounds with visual query interactions between each, APE systematically elicited the participant's intent. Following the first generation (I1), core stylistic and structural decisions were established, including a modern minimalist style, a minimalist geometric four-poster bed frame, natural wood flooring with warm tones, nature-inspired large-scale wall murals, integrated ambient LED lighting, a fireplace with a light natural stone surround, and a flat ceiling with integrated cove lighting. Between I2 and I3, the aesthetic direction was further articulated by specifying soft pastel wall color palettes contrasted with contemporary pop art pieces featuring abstract shapes and bright colors, as well as by coordinating the bed frame material and color with the mural palette and introducing glossy white lacquered furnishings. Subsequent rounds refined material and furnishing details: window treatments and furniture material and style (I4); ceiling finish and lighting (I5); and side table material (I6).

This trace establishes broad aesthetic and compositional foundations early, then progressively adds material specifications and detail-level choices, exemplifying a hierarchical refinement pattern observed across APE sessions: The final prompt (160 words) is 8× more detailed than the initial specification, yet achieved in only six generation rounds through structured elicitation. The participant's selections reveal preference discovery—elements like

Original Prompt: A large bedroom with open spaces, and that includes a fire place, four-poster bed, big windows and lots of natural light.

Select Answer Option
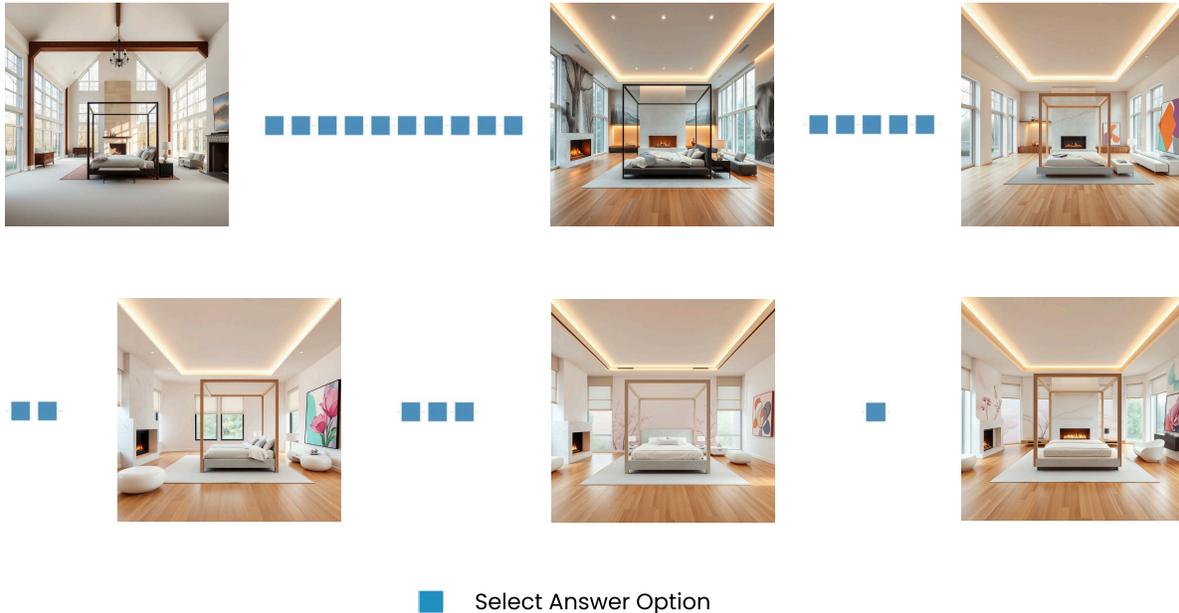
Final Prompt: A large, open-plan contemporary bedroom featuring a modern minimalist fireplace with a light natural stone surround and hearth in marble or limestone. The centerpiece is a four-poster bed with a minimalist geometric frame made of light pastel-painted wood echoing the soft pastel wall mural colors of light pinks, lavenders, and mint greens. Multiple large windows with minimalist neutral roller shades fill the space with abundant natural light, complemented by subtle under-furniture accent strip lighting beneath the bed and side tables. Flooring is warm-toned natural wood fully exposed, with no rugs. Walls showcase large-scale nature-inspired murals in soft pastels alongside bold pop art wall paintings or prints of abstract shapes and bright colors. Furnishings include glossy white lacquered molded plastic or resin furniture with smooth curves, paired with simple-based side tables or nightstands topped with natural stone like marble or limestone. The ceiling is flat and smooth with a matte soft neutral paint finish and integrated cove lighting for ambient illumination.

Figure 9: Interaction trace from an APE participant in an interior design task. Beginning from a brief functional prompt, the participant's intent is progressively refined across six generation rounds (I1–I6). Generated images are ordered temporally, and blue squares indicate visual query interactions. Early rounds establish high-level stylistic and compositional choices (e.g., modern minimalist style, spatial layout, and major architectural elements), followed by mid-stage specification of aesthetic direction (color palette, art style, and material coherence), and later-stage refinement of fine-grained details (window treatments, lighting configuration, and furniture materials), exemplifying APE's hierarchical elicitation process and visually grounded intent formation.

the pastel-to-pop-art color contrast, stone fireplace surround, and under-furniture LED strips emerged through visual comparison rather than initial articulation, illustrating how APE facilitates intent formation beyond mere preference elicitation.

5.4.2 Query Effectiveness. To assess the quality of system-generated queries, we measured how often participants found them useful. On average, 53.19% ($SE = 2.71\%$) of questions were reported as useful by the participants. While this represents roughly half of all queries, it is important to note that this metric reflects explicit

usefulness ratings; queries that did not receive positive ratings may still have helped narrow the design space or clarified user preferences implicitly. This finding highlights both the effectiveness of the information-theoretic query selection strategy and the highly personalized nature of user intent.

Qualitative feedback supported this interpretation. Participants commented that APE's visual queries were generally effective at eliciting discriminative preferences, such as motifs, styles, and composition, quickly narrowing the design space: "*The visual prompts allowed me to quickly get the idea of the kind of image I wanted to generate.*" Nonetheless, the effectiveness of queries depended on the coverage and granularity of the offered options; a few users requested more variety or noted that certain fine details could not be captured through the available options.

5.4.3 Iteration Requirements. We tracked how many generation attempts participants required before achieving satisfactory results. APE users needed significantly fewer iterations of image generation ($M = 4.88, SE = 0.41$) compared to Baseline users ($M = 8.06, SE = 0.85; U = 2675.0, p = 0.003, r = -0.31$), corresponding to a 39.5% reduction in iteration cost. Qualitative feedback supported this efficiency gain, with one APE participant noting: "*Quite fast I found the perfect image, very close to what I imagined.*"

However, this advantage may be partially offset by increased latency due to additional API calls and background computation in APE. Our study employed FLUX.1, a flow-matching-based generative model for rapid inference. One Baseline participant with prior experience using slower systems (Midjourney, DALL·E) commented: "*I was impressed with how fast the images were generated,*" suggesting FLUX's speed may diminish the perceived value of iteration reduction. Conversely, APE's additional computational overhead drew occasional criticism: "*The system could be a little faster.*" The efficiency trade-off likely shifts considerably with slower base models. For text-to-image systems requiring 30-60 seconds per generation, reducing iterations from 8 to 5 would save 1.5-3 minutes of cumulative wait time—a more substantial practical benefit. Future implementations could employ hybrid strategies: using fast distilled models (e.g., FLUX.1-schnell) for generating visual query exemplars during elicitation, then applying the refined prompt to slower, higher-quality models for final generation. This approach would preserve APE's interaction efficiency while accommodating diverse model speed-quality trade-offs.

6 Discussion

The central challenge in text-to-image generation is not merely model capacity but the *alignment gap*: the mismatch between users' latent visual intentions and their ability to articulate these intentions through text. Our work provides evidence that reframing prompting as bidirectional communication, where systems actively elicit preferences through strategic visual queries, can significantly reduce this gap.

Across controlled technical evaluation (Section 4) and real-world user studies (Section 5), a consistent pattern emerges: information-theoretically optimized visual queries enable users to specify intent with greater precision than traditional prompt engineering, while requiring fewer iterations and imposing no additional cognitive burden. These benefits stem not from bypassing linguistic articulation,

but from scaffolding it—transforming the task from open-ended prompt construction to guided refinement through comparative judgment. We now examine: (1) why visual query interaction proves effective for preference elicitation, (2) how transparent representations enhance user agency, (3) the tension between convergent elicitation and exploratory creativity, and (4) limitations and future directions.

6.1 Visual Query Interaction for Intent Elicitation

Our results validate a core design principle underlying APE: that visual query interaction provides an effective complement to textual articulation for preference elicitation. By presenting visual options that embody specific feature values, APE transforms preference specification from open-ended articulation into comparative judgment, directly addressing documented difficulties in verbally specifying rich visual concepts [54, 69] by leveraging the cognitive advantage of recognition over generation [70]. The significant improvements across all generative alignment dimensions suggest that this two-way communication enhances perceived alignment [38, 72].

The reduction in image generation iterations represents substantial practical value in real-world use cases, consistent with prior work showing interactive query-based prompt optimization reduces search costs in image generation [27, 82]. A generation cycle involves cognitive effort (evaluating results, deciding on modifications), time (waiting for generation), and computational cost (running the text-to-image model). By nearly halving the required iterations, APE transforms user experience from exhausting trial-and-error to structured refinement. This efficiency improvement appears particularly valuable for users with clear conceptual goals but limited prompt engineering strategies, though our exploratory subgroup analysis was underpowered for definitive claims about experience-based moderation.

6.2 Transparency and User Agency in AI-Assisted Creation

Besides query-based interaction, an important benefit emerged from APE's explicit feature requirement representation: transparency that supports both understanding and control. Participants commented positively on seeing preferences as structured requirements, with one noting: "*It was innovative and smart that I could see the filters that are being applied before creating an image and being able to edit them. In this way, I can understand that the AI has understood what I am asking for.*" This transparency likely contributed to APE's particularly strong improvement in the instruction gap dimension, where users reported better ability to provide appropriate input. The editable requirements panel also enabled mid-process corrections—users could identify and revise specific features rather than restarting from scratch—which may explain the reduced intentionality gap as users gained confidence that their vision was accurately captured [38]. This contrasts with black-box prompting, where users input text without understanding model interpretation, limiting opportunities for bias detection, error correction, or trust-building through interpretability.

The mixed-initiative design [16] of APE also reflects emerging principles for human-AI collaboration. In APE, the structured requirements panel serves as an explainable interface helping user to identify when the system encodes biased assumptions or makes incorrect inferences about their intent [80]. The text editing capability then provides *user agency* to correct these errors, consistent with research showing that editable AI outputs better support user-driven value alignment than uneditable suggestions [19]. Notably, these safeguards require users to recognize biases, which may not occur if biases align with users' own assumptions or operate subtly [2, 8, 26]. Future work could explore proactive such as diverse sampling or bias detection like OpenBias [17] to further strengthen these protections.

6.3 Creative Exploration Through Structured Elicitation

While APE optimizes for efficient convergence to user intent, an interesting pattern emerged in user feedback: some participants reported that structured queries facilitated creative exploration. Multiple participants noted discovering preferences through visual options they had not initially considered: "*The question about lighting made me realize I cared about that. I hadn't thought about it before, but seeing the options, I definitely had a preference.*" and "*The system questions helped in getting the creativity flowing.*" This suggests that APE's queries may serve dual purposes: efficiently gathering preference information while scaffolding preference discovery through systematic consideration of visual dimensions users might otherwise overlook. This finding connects to research on creativity support tools showing that constraints can enhance rather than hinder creativity by providing structure for exploration [76]. APE's queries function as design prompts that encourage systematic thinking about visual dimensions, potentially supporting creative development alongside efficient elicitation.

However, we note important scope limitations: while we observed anecdotal evidence of preference discovery, APE's core design is grounded in traditional preference elicitation frameworks that assume users have pre-existing preferences to uncover [12, 24, 81]. Our information-theoretic query strategy explicitly targets efficient convergence to latent user intent, making APE well-suited for *goal-directed tasks* where users face articulation barriers. For purely *exploratory tasks* prioritizing serendipitous discovery, alternative approaches emphasizing divergent exploration over convergent elicitation may prove more appropriate. The tension between efficient preference elicitation and open-ended creative exploration represents an important theoretical distinction that future work should address through adaptive interaction modes.

6.4 Limitations and Future Work

We identify several important limitations that bound the applicability of our approach and suggest directions for future work.

Firstly, APE addresses misalignment by reducing epistemic uncertainty (the gap between user intent and system understanding) but does not address inherent limitations of image generation models (aleatoric uncertainty). The method's effectiveness is fundamentally bounded by model capacity: as generative models improve, APE's potential increases; conversely, with limited models, alignment

gains may be constrained. Future work could explore synergistic approaches that integrate APE with model improvement techniques, such as user-specific fine-tuning [91] or self-improvement via consistency regularization [56], to jointly enhance alignment and output quality.

Secondly, our information-theoretic selection optimizes for efficiency, assuming users have clear (though unarticulated) preferences and a proper model of user intention. In reality, preferences often form dynamically and correlatively. A few participants noted the desire for additional flexibility, such as the ability to skip or undo queries, or found some question sequences somewhat lengthy: "*I wish it were easier to go back, though, or to skip a question.*" Future interfaces might provide more support for calibrating misspecified priors, and balance efficiency with exploration support, occasionally selecting queries that maximize creative insight rather than immediate information gain, though designing such balanced objectives remains an open challenge.

Thirdly, APE's effectiveness depends critically on LLM quality for feature generation, completion sampling, and prompt synthesis. The LLM persona may encode biases from pre-training data, such as demographic, cultural, or aesthetic biases [2, 8, 26, 65]. While our mixed-initiative design provides mitigation mechanisms (transparent panel for bias detection [80], editing for user agency [19]), these require users to recognize biases, which may not occur if biases align with users' assumptions or operate subtly. Future work could investigate proactive bias detection, diverse sampling strategies [17], and diversity-aware query generation.

Lastly, while the model-agnostic design of APE enables its application to a broad range of text-to-image generation models, our empirical evaluation is limited to a single state-of-the-art model. Future work could explore its generalizability across a wider set of generative architectures to assess robustness and uncover model-specific factors. In addition, investigating APE's potential to extend across modalities (e.g. text-to-music [14] and text-to-video [74] generation) offers a compelling avenue for further study. As generative models continue to advance, addressing the alignment gap becomes increasingly critical beyond improvements in model capacity alone. APE constitutes an initial step toward mitigating this gap through principled interaction design, pointing toward future creative AI systems that are more interactive, transparent, and adaptive than current text-based interfaces.

7 Conclusion

Aligning text-to-image generation with user intent remains a persistent challenge when users struggle to articulate rich visual concepts through text. We presented ADAPTIVE PROMPT ELICITATION (APE), which reframes prompting as bidirectional communication through information-theoretically optimized visual queries. By eliciting preferences through strategic visual comparisons, APE transforms intent specification from unilateral articulation into guided interactive refinement. Evaluation across controlled benchmarks and user studies demonstrates that this interactive approach achieves stronger alignment with improved efficiency compared to manual prompting and automatic optimization. As generative models advance, principled interaction design that actively elicits user intent offers

a complementary path toward alignment alongside improvements in model capacity.

Acknowledgments

References

[1] Krzysztof Adamkiewicz, Paweł W. Woźniak, Julia Dominiak, Andrzej Romanowski, Jakob Karolus, and Stanislav Frolov. 2025. PromptMap: An Alternative Interaction Style for AI-Based Image Generation. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 1162–1176. doi:10.1145/3708359.3712150

[2] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1493–1504. doi:10.1145/3593013.3594095

[3] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. 2024. Training Diffusion Models with Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*.

[4] 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 (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3586183.3606725

[5] Eric Brochu, Tyson Brochu, and Nando de Freitas. 2010. A Bayesian interactive optimization approach to procedural animation design. In *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 103–112.

[6] Chengkun Cai, Haoliang Liu, Xu Zhao, Zhongyu Jiang, Tianfang Zhang, Zongkai Wu, John Lee, Jenq-Neng Hwang, and Lei Li. 2025. Bayesian Optimization for Controlled Image Editing via LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, Vienna, Austria, 10045–10056. doi:10.18653/v1/2025.findings-acl.523

[7] Tingfeng Cao, Chengyu Wang, Bingyan Liu, Ziheng Wu, Jinhui Zhu, and Jun Huang. 2023. BeautifulPrompt: Towards Automatic Prompt Engineering for Text-to-Image Synthesis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Association for Computational Linguistics, Singapore, 1–11. doi:10.18653/v1/2023.emnlp-industry.1

[8] Jane Castleman and Aleksandra Korolova. 2025. Adultification Bias in LLMs and Text-to-Image Models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 2751–2767. doi:10.1145/3715275.3732178

[9] Kathryn Chaloner and Isabella Verdinelli. 1995. Bayesian Experimental Design: A Review. *Statist. Sci.* 10, 3 (1995), 273–304.

[10] Zijie Chen, Lichao Zhang, Fangsheng Weng, Lili Pan, and Zhenzhong Lan. 2024. Tailored Visions: Enhancing Text-to-Image Generation with Personalized Prompt Rewriting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7727–7736.

[11] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. *Advances in Neural Information Processing Systems* 30 (2017).

[12] Wei Chu and Zoubin Ghahramani. 2005. Preference Learning with Gaussian Processes. In *Proceedings of the 22nd International Conference on Machine Learning - ICML '05*. ACM Press, Bonn, Germany, 137–144. doi:10.1145/1102351.1102369

[13] John Joon Young Chung and Eytan Adar. 2023. PromptPaint: Steering Text-to-Image Generation Through Paint Medium-like Interactions. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3586183.3606777

[14] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Defossez. 2023. Simple and Controllable Music Generation. *Advances in Neural Information Processing Systems* 36 (2023), 47704–47720.

[15] Siddhartha Datta, Alexander Ku, Deepak Ramachandran, and Peter Anderson. 2024. Prompt Expansion for Adaptive Text-to-Image Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 3449–3476. doi:10.18653/v1/2024.acl-long.189

[16] Sebastian Deterding, Jonathan Hook, Rebecca Fiebrink, Marco Gillies, Jeremy Gow, Memo Akten, Gillian Smith, Antonios Liapis, and Kate Compton. 2017. Mixed-Initiative Creative Interfaces. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 628–635. doi:10.1145/3027063.3027072

[17] Moreno D'Incà, Elia Peruzzo, Massimiliano Mancini, Dejia Xu, Vidit Goe, Xingqian Xu, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. 2024. OpenBias: Open-Set Bias Detection in Text-to-Image Generative Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12225–12235. doi:10.1109/CVPR52733.2024.01162

[18] Jerry Alan Fails and Dan R. Olsen. 2003. Interactive Machine Learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces (IUI '03)*. Association for Computing Machinery, New York, NY, USA, 39–45. doi:10.1145/604045.604056

[19] Xianzhe Fan, Qing Xiao, Xuhui Zhou, Jiaxin Pei, Maarten Sap, Zhicong Lu, and Hong Shen. 2025. User-Driven Value Alignment: Understanding Users' Perceptions and Strategies for Addressing Biased and Discriminatory Statements in AI Companions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3706598.3713477

[20] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160.

[21] 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

[22] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. 2023. DreamSim: Learning New Dimensions of Human Visual Similarity Using Synthetic Data. *Advances in Neural Information Processing Systems* 36 (2023), 50742–50768.

[23] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep Bayesian Active Learning with Image Data. In *Proceedings of the 34th International Conference on Machine Learning*. PMLR, 1183–1192.

[24] Javier González, Zhenwen Dai, Andreas Damianou, and Neil D. Lawrence. 2017. Preferential Bayesian Optimization. In *Proceedings of the 34th International Conference on Machine Learning*. PMLR, 1282–1291.

[25] Yuhan Guo, Hanning Shao, Can Liu, Kai Xu, and Xiaoru Yuan. 2025. PrompTHis: Visualizing the Process and Influence of Prompt Editing during Text-to-Image Creation. 31, 9 (2025), 4547–4559. doi:10.1109/TVCG.2024.3408255

[26] Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. In *The Twelfth International Conference on Learning Representations*.

[27] Meera Hahn, Wenjun Zeng, Nithish Kannen, Rich Galt, Kartikeya Badola, Been Kim, and Zi Wang. 2025. Proactive Agents for Multi-Turn Text-to-Image Generation Under Uncertainty. In *Forty-Second International Conference on Machine Learning*.

[28] Shu-Jung Han and Susan R. Fussell. 2025. Understanding User Perceptions and the Role of AI Image Generators in Image Creation Workflows. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2025-04-25) *(CHI '25)*. Association for Computing Machinery, 1–17. doi:10.1145/3706598.3713227

[29] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2023. Optimizing Prompts for Text-to-Image Generation. *Advances in Neural Information Processing Systems* 36 (2023), 66923–66939.

[30] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[31] Yutong He, Alexander Robey, Naoki Murata, Yiding Jiang, Joshua Nathaniel Williams, George J. Pappas, Hamed Hassani, Yuki Mitsufuji, Ruslan Salakhutdinov, and J Zico Kolter. 2025. Automated Black-box Prompt Engineering for Personalized Text-to-Image Generation. *Transactions on Machine Learning Research* (2025).

[32] Nailei Hei, Qianyu Guo, Zihao Wang, Yan Wang, Haofen Wang, and Wenqiang Zhang. 2024. A User-Friendly Framework for Generating Model-Preferred Prompts in Text-to-Image Synthesis. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 3 (March 2024), 2139–2147. doi:10.1609/aaai.v38i3.27986

[33] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian Active Learning for Classification and Preference Learning. arXiv:1112.5745 [stat.ML] https://arxiv.org/abs/1112.5745

[34] Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative health research* 15, 9 (2005), 1277–1288.

[35] Daolang Huang, Yujia Guo, Luigi Acerbi, and Samuel Kaski. 2024. Amortized Bayesian Experimental Design for Decision-Making. *Advances in Neural Information Processing Systems* 37 (2024), 109460–109486.

[36] Rahul Jain, Amit Goel, Koichiro Niinuma, and Aakar Gupta. 2025. AdaptiveSliders: User-aligned Semantic Slider-based Editing of Text-to-Image Model Output. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–27. doi:10.1145/3706598.3714292

[37] Kaixuan Ji, Jiafan He, and Quanquan Gu. 2025. Reinforcement Learning from Human Feedback with Active Queries. *Transactions on Machine Learning Research* (2025).

[38] Soomin Kim, Jinsu Eun, Changhoon Oh, and Joonhwan Lee. 2025. "Journey of Finding the Best Query": Understanding the User Experience of AI Image Generation System. *International Journal of Human–Computer Interaction* 41, 2 (2025), 951–969. doi:10.1080/10447318.2024.2307670

[39] Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. 2019. BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning. *Advances in Neural Information Processing Systems* 32 (2019).

[40] Kasia Kobalczyk, Nicolás Astorga, Tennison Liu, and Mihaela van der Schaar. 2025. Active Task Disambiguation with LLMs. In *The Thirteenth International Conference on Learning Representations*.

[41] Yuki Koyama and Masataka Goto. 2022. BO as Assistant: Using Bayesian Optimization for Asynchronously Generating Design Suggestions. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3526113.3545664

[42] Yuki Koyama, Issei Sato, and Masataka Goto. 2020. Sequential Gallery for Interactive Visual Design Optimization. *ACM Trans. Graph.* 39, 4 (Aug. 2020), 88:88:1–88:88:12. doi:10.1145/3386569.3392444

[43] Black Forest Labs. 2025. FLUX.1 - a Black-Forest-Labs Collection. https://huggingface.co/collections/black-forest-labs/flux1 Accessed: 2025-10-10.

[44] Daniel Lee, Nikhil Sharma, Donghoon Shin, DaEun Choi, Harsh Sharma, Jeonghwan Kim, and Heng Ji. 2025. ThematicPlane: Bridging Tacit User Intent and Latent Spaces for Image Generation. In *Adjunct Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST Adjunct '25)*. Association for Computing Machinery, New York, NY, USA, Article 120, 3 pages. doi:10.1145/3746058.3758376

[45] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. 2023. Aligning Text-to-Image Models using Human Feedback. arXiv:2302.12192 [cs.LG] https://arxiv.org/abs/2302.12192

[46] Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. 2025. Eliciting Human Preferences with Language Models. In *The Thirteenth International Conference on Learning Representations*.

[47] Chen Liang, Lianghua Huang, Jingwu Fang, Huanzhang Dou, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Junge Zhang, Xin Zhao, and Yu Liu. 2025. IDEA-Bench: How Far Are Generative Models from Professional Designing?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18541–18551.

[48] Haichuan Lin, Yilin Ye, Jiazhi Xia, and Wei Zeng. 2025. SketchFlex: Facilitating Spatial-Semantic Coherence in Text-to-Image Generation with Region-Based Sketches. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3706598.3713801

[49] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating Text-to-Visual Generation with Image-to-Text Generation. In *European Conference on Computer Vision*. Springer, 366–384.

[50] Shihong Liu, Samuel Yu, Zhiqiu Lin, Deepak Pathak, and Deva Ramanan. 2024. Language Models as Black-Box Optimizers for Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12687–12697.

[51] Vivian Liu and Lydia B Chilton. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2022-04-29) (CHI '22). Association for Computing Machinery, 1–23. doi:10.1145/3491102.3501825

[52] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J. Cai. 2020. Novice-AI Music Co-Creation via AI-Steering Tools for Deep Generative Models. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376739

[53] Shuai Ma, Zijun Wei, Feng Tian, Xiangmin Fan, Jianming Zhang, Xiaohui Shen, Zhe Lin, Jin Huang, Radomír Měch, Dimitris Samaras, et al. 2019. SmartEye: assisting instant photo taking via integrating user preference with deep view proposal network. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[54] Atefeh Mahdavi Goloujeh, Anne Sullivan, and Brian Magerko. 2024. Is It AI or Is It Me? Understanding Users' Prompt Journey with Text-to-Image Generative AI Tools. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing*

[55] Saaduddin Mahmud, Mason Nakamura, and Shlomo Zilberstein. 2025. MAPLE: A Framework for Active Preference Learning Guided by Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 26 (April 2025), 27518–27528. doi:10.1609/aaai.v39i26.34964

[56] Oscar Mañas, Pietro Astolfi, Melissa Hall, Candace Ross, Jack Urbanek, Adina Williams, Aishwarya Agrawal, Adriana Romero-Soriano, and Michal Drozdzal. 2024. Improving Text-to-Image Consistency via Automatic Prompt Optimization. *Transactions on Machine Learning Research* (2024).

[57] Henry B Mann and Donald R Whitney. 1947. On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* 18, 1 (1947), 50–60. jstor:2236101

[58] OpenAI. 2025. Vector embeddings. https://platform.openai.com/docs/guides/embeddings Accessed: 2025-10-10.

[59] Jonas Oppenlaender. 2024. A Taxonomy of Prompt Modifiers for Text-to-Image Generation. *Behaviour & Information Technology* 43, 15 (Nov. 2024), 3763–3776. doi:10.1080/0144929X.2023.2286532

[60] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research* (2024).

[61] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[62] Srishti Palani and Gonzalo Ramos. 2024. Evolving Roles and Workflows of Creative Practitioners in the Age of Generative AI. In *Proceedings of the 16th Conference on Creativity & Cognition (C&C '24)*. Association for Computing Machinery, New York, NY, USA, 170–184. doi:10.1145/3635636.3656190

[63] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model Is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, Vol. 36. 53728–53741.

[64] Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. 2024. Modern Bayesian Experimental Design. *Statist. Sci.* 39, 1 (2024), 100–114.

[65] Philip Resnik. 2025. Large Language Models Are Biased Because They Are Large Language Models. *Computational Linguistics* 51, 3 (Sept. 2025), 885–906. doi:10.1162/coli_a_00558

[66] Jeba Rezwana and Mary Lou Maher. 2023. Designing Creative AI Partners with COFI: A Framework for Modeling Interaction in Human-AI Co-Creative Systems. *ACM Trans. Comput.-Hum. Interact.* 30, 5, Article 67 (Sept. 2023), 28 pages. doi:10.1145/3519026

[67] 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.

[68] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494.

[69] Téo Sanchez. 2023. Examining the Text-to-Image Community of Practice: Why and How Do People Prompt Generative AIs?. In *Proceedings of the 15th Conference on Creativity and Cognition* (New York, NY, USA, 2023-06-19) (C&C '23). Association for Computing Machinery, 43–61. doi:10.1145/3591196.3593051

[70] Jonathan W Schooler and Tonya Y Engstler-Schooler. 1990. Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology* 22, 1 (1990), 36–71. doi:10.1016/0010-0285(90)90003-M

[71] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347 [cs.LG] https://arxiv.org/abs/1707.06347

[72] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary Chase Lipton, Qiaozhu Mei, Michael Terry, Diyi Yang, Meredith Ringel Morris, Paul Resnick, and David Jurgens. 2025. Position: Towards Bidirectional Human-AI Alignment. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*.

[73] 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,

1–20. doi:10.1145/3706598.3714087

[74] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oran Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. 2023. Make-A-Video: Text-to-Video Generation without Text-Video Data. In *The Eleventh International Conference on Learning Representations*.

[75] Kihoon Son, Kyungmin Kim, and Kyung Hoon Hyun. 2022. BIGexplore: Bayesian Information Gain Framework for Information Exploration. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3491102.3517729

[76] Patricia D Stokes. 2005. *Creativity from Constraints: The Psychology of Breakthrough*. Springer Publishing Company.

[77] 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 (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3613904.3642754

[78] Yuki Tatsukawa, I-Chao Shen, Mustafa Doga Dogan, Anran Qi, Yuki Koyama, Ariel Shamir, and Takeo Igarashi. 2025. FontCraft: Multimodal Font Design Using Interactive Bayesian Optimization. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3706598.3713863

[79] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. 2024. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. *Advances in Neural Information Processing Systems* 37 (2024), 84839–84865.

[80] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1 (April 2023), 129:1–129:38. doi:10.1145/3579605

[81] Paolo Viappiani. 2014. Preference Modeling and Preference Elicitation: An Overview. In *Proceedings of the First International Workshop on Decision Making and Recommender Systems (DMRS2014)*, Vol. 1278. CEUR Workshop Proceedings, Bolzano, Italy, 19–24.

[82] Jianhui Wang, Yangfan He, Yan Zhong, Xinyuan Song, Jiayi Su, Yuheng Feng, Ruoyu Wang, Hongyang He, Wenyu Zhu, Xinhang Yuan, Miao Zhang, Keqin Li, Jiaqi Chen, Tianyu Shi, and Xueqian Wang. 2025. Twin Co-Adaptive Dialogue for Progressive Image Generation. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*. Association for Computing Machinery, New York, NY, USA, 3645–3653. doi:10.1145/3746027.3755141

[83] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. arXiv:2402.05672 [cs.CL] https://arxiv.org/abs/2402.05672

[84] 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 2024 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2024-05-11) *(CHI '24)*. Association for Computing Machinery, 1–21. doi:10.1145/3613904.3642803

[85] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2023. DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 893–911. doi:10.18653/v1/2023.acl-long.51

[86] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[87] 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 (WWW '23)*. Association for Computing Machinery, New York, NY, USA, 3892–3902. doi:10.1145/3543507.3587430

[88] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. *Advances in Neural Information Processing Systems* 36 (2023), 15903–15935.

[89] Kenta Yamamoto, Yuki Koyama, and Yoichi Ochiai. 2022. Photographic Lighting Design with Photographer-in-the-Loop Bayesian Optimization. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3526113.3545690

[90] 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* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. doi:10.1145/3544548.3581388

[91] Xingchen Zeng, Ziyao Gao, Yilin Ye, and Wei Zeng. 2024. IntentTuner: An Interactive Framework for Integrating Human Intentions in Fine-tuning Text-to-Image Generative Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3613904.3642165

[92] Mingxu Zhou, Dengming Zhang, Weitao You, Ziqi Yu, Yifei Wu, Chenghao Pan, Huiting Liu, Tianyu Lao, and Pei Chen. 2024. StyleFactory: Towards Better Style Alignment in Image Creation through Style-Strength-Based Control and Evaluation. 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 107, 15 pages. doi:10.1145/3654777.3676370

A User Study Details

A.1 Design Tasks

The eight text-to-image generation scenarios used in the user study are summarized in Table 3.
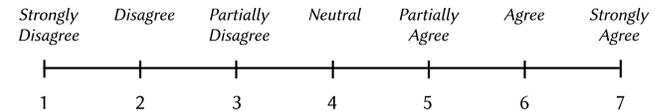
A.2 Pre-Study Questionnaire

Before starting the design tasks, participants completed a pre-study questionnaire to gather information about their background and experience with AI tools and design activities. The participants' response are summarized in Table 4.

A.3 Post-Study Questionnaire

In this section, we detail the design of post-study questionnaires and participants' response.

A.3.1 Alignment Questionnaire. Participants rated their agreement with the following statements on a 7-point Likert scale:

Strongly Disagree	Disagree	Partially Disagree	Neutral	Partially Agree	Agree	Strongly Agree
1	2	3	4	5	6	7

Capability Gap

+ "The system supported me in working out how to generate the best possible image based on what I wanted."
− "The system led me away from what I wanted, and the image did not match what I had aimed for."

Instruction Gap

+ "The system helped me provide the right input to generate the image as I intended."
− "I was disappointed by how the system interpreted my input when generating the image."

Intentionality Gap

+ "The system helped me clarify what I wanted from the image."
− "I was unsure how (or if) the system could generate an image more in line with my intention."

Envision Overall

+ "The system made me think of a wider range of possible images that fit my goal."
− "The system made it harder for me to imagine alternative images that might better fit my goal."

Table 3: Text-to-image generation scenarios used in the user study.

Scenario	Description
Interior Design	Imagine and create an image of your ideal room.
Painting	Visualize your dream travel destination as a painting.
Photography	Imagine and generate a photo of a dish or drink you would love to have right now.
App Icon	Create an icon for an app you use or would like to invent.
Poster Design	Design a poster for an event or celebration.
Logo Design	Design a logo for a project, team, or group.
Fashion Design	Imagine and design an outfit you would like to wear.
Architectural Style	Visualize a building you would like to visit.

Table 4: Frequency of AI tool usage and design activities.

Theme	Question	Usage Frequency	Count
LLM Usage	How often do you use large language models (LLMs) such as ChatGPT, Claude, or DeepSeek?	Never	2
		Less than once a month	3
		1–3 times per month	11
		1–6 times per week	42
		Daily or almost daily	70
Text-to-Image Generation	What is your experience with text-to-image generation (e.g., Midjourney, Stable Diffusion, DALL·E, FLUX)?	Never	23
		I've tried once or twice	34
		1–3 times per month	36
		1–6 times per week	26
		Daily or almost daily	9
Design Activities	How often do you engage in design-related activities (graphic design, visual arts, UX/UI)?	Never	21
		Once a year or less	24
		A few times a year	31
		Weekly	15
		Monthly	25
		Daily or almost daily	12

A.3.2 Post-Study Questionnaire Results. Figure 10 shows the distribution of user ratings in the alignment questionnaire. Figure 11 shows the distribution of user ratings in the NASA-TLX questionnaire.

B Generated Images

B.1 Technical Evaluation

Some examples of resulting images in the technical evaluation are shown in Figure 12.

B.2 User Evaluation

Examples of generated images from participants in APE and Baseline conditions are shown in Figure 13 and Figure 14, respectively.
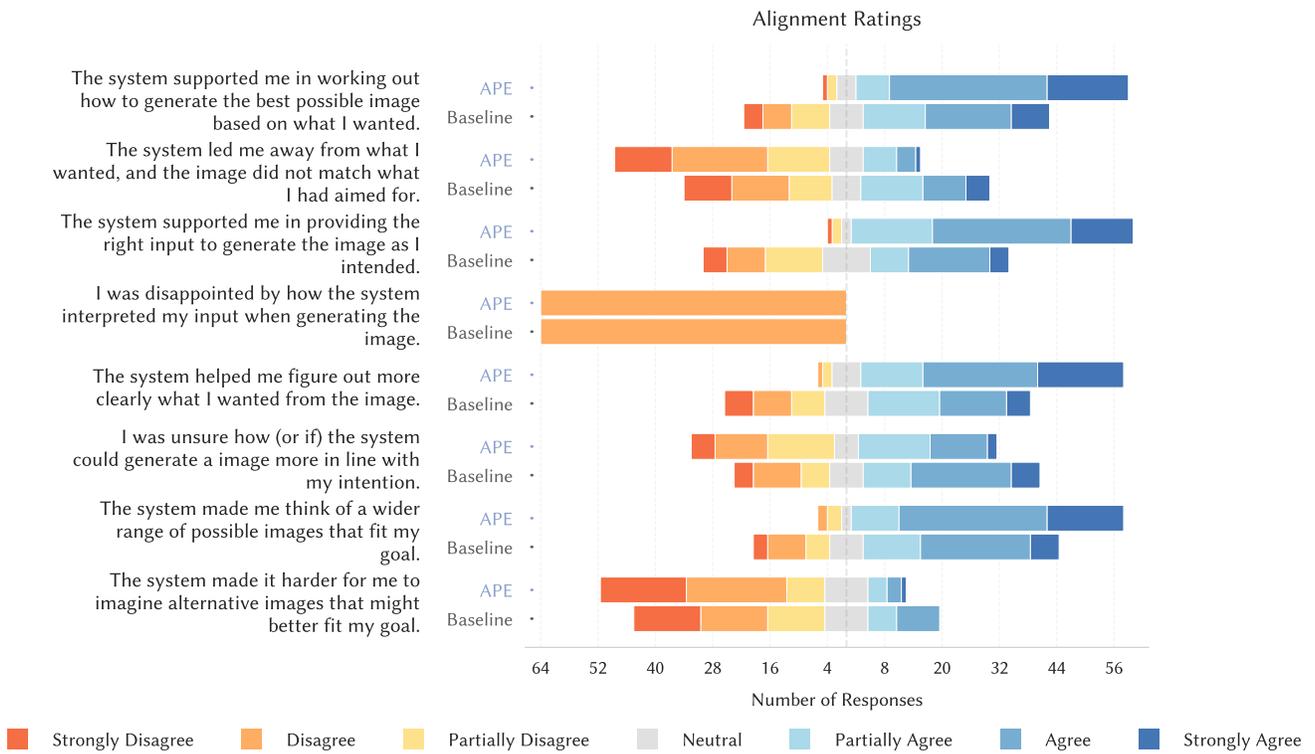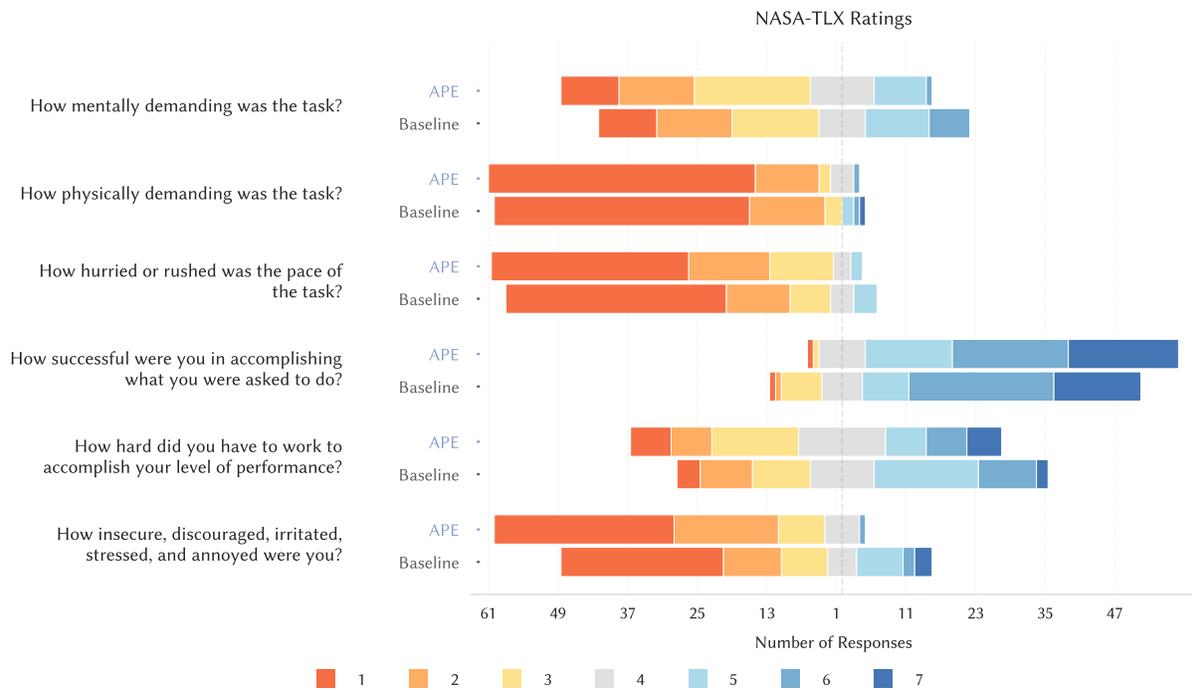
Alignment Ratings

The system supported me in working out how to generate the best possible image based on what I wanted.
- APE
- Baseline

The system led me away from what I wanted, and the image did not match what I had aimed for.
- APE
- Baseline

The system supported me in providing the right input to generate the image as I intended.
- APE
- Baseline

I was disappointed by how the system interpreted my input when generating the image.
- APE
- Baseline

The system helped me figure out more clearly what I wanted from the image.
- APE
- Baseline

I was unsure how (or if) the system could generate a image more in line with my intention.
- APE
- Baseline

The system made me think of a wider range of possible images that fit my goal.
- APE
- Baseline

The system made it harder for me to imagine alternative images that might better fit my goal.
- APE
- Baseline

Number of Responses

Legend: Strongly Disagree | Disagree | Partially Disagree | Neutral | Partially Agree | Agree | Strongly Agree

Figure 10: Alignment ratings by participants

NASA-TLX Ratings

How mentally demanding was the task?
- APE
- Baseline

How physically demanding was the task?
- APE
- Baseline

How hurried or rushed was the pace of the task?
- APE
- Baseline

How successful were you in accomplishing what you were asked to do?
- APE
- Baseline

How hard did you have to work to accomplish your level of performance?
- APE
- Baseline

How insecure, discouraged, irritated, stressed, and annoyed were you?
- APE
- Baseline

Number of Responses

Legend: 1 | 2 | 3 | 4 | 5 | 6 | 7

Figure 11: NASA TLX ratings by participants

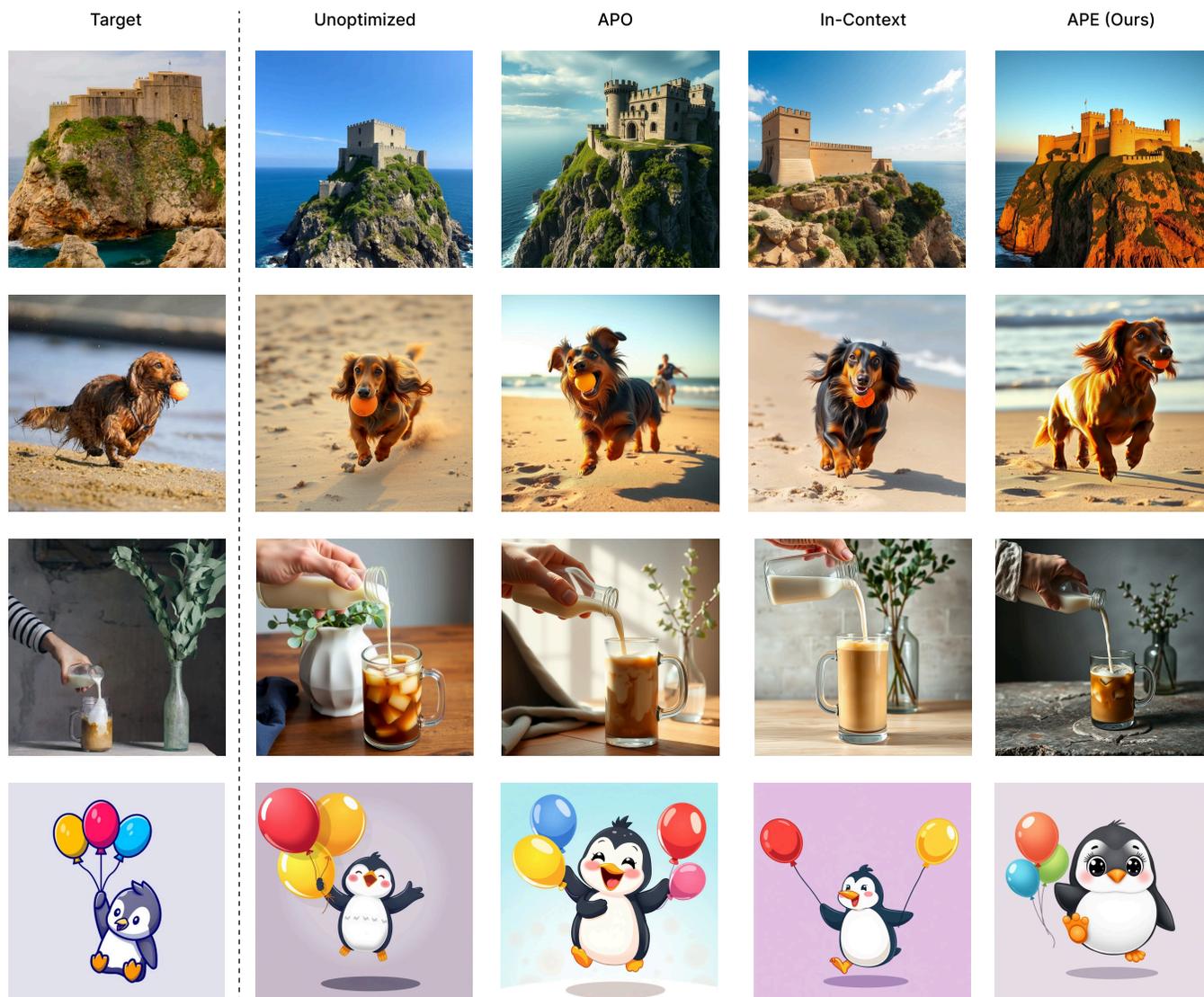| Target	Unoptimized	APO	In-Context	APE (Ours)

Figure 12: Examples of resulting images across methods in the technical evaluation. For each row, we show the target reference image (left) followed by outputs from four approaches: Unoptimized, APO, In-Context Query, and APE (our method). APE consistently produces images with stronger visual alignment to target references, capturing key features such as composition and color palette (row 1), perspective and motion dynamics (row 2), tonal atmosphere and spatial arrangement (row 3), and pose (row 4).

Original Prompt: produce an image in painting style of the mountains in new zealand, with a glacier at the base, snow capped peaks, and brown grass at the base, make it look like something out of Lord of the Rings
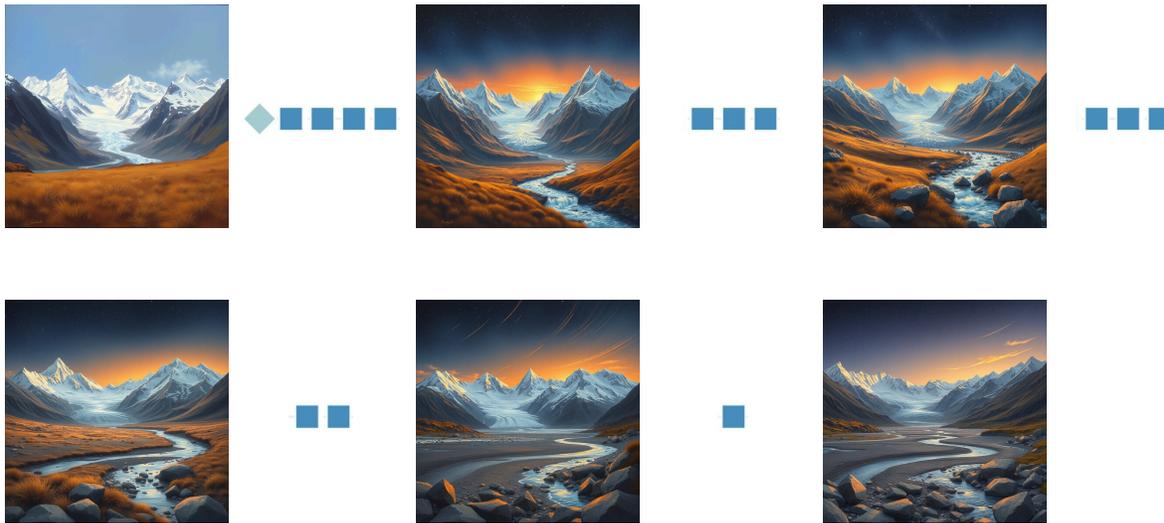

■ Select Answer Option   ◆ Provide Other Answer

Final Prompt: A detailed painting of New Zealand mountains featuring a mix of sharp snow-capped peaks and gentle slopes, with a glacier base partially covered in snow, scattered rocks and debris, and exposed rocky soil with minimal vegetation. A meandering river with gravel bars and pools flows through the scene, while smooth, polished rocks and boulders occupy the foreground. The atmosphere is epic and fantasy-inspired, reminiscent of Lord of the Rings, illuminated by warm sunrise lighting. The sky transitions realistically from a starry night with detailed stars to dawn colors in cool, muted tones of lavender, pale yellow, and soft grey, accented by thin wispy cirrus clouds catching the sunrise light. The overall color palette consists of muted earthy tones with subtle warm highlights, focusing solely on landscape without any visible wildlife or creatures.

Figure 13: Examples of generated images from an APE participant. The original prompt (top) provides a general theme; the user progressively clarifies their intent through visual queries (square: selected options; diamond: custom answer) between generations (I1–I6). Requirements accumulate from foundational elements (I2: sky, mountains, water) to compositional details (I3: vegetation, foreground) to aesthetic refinements and fine-grained features (I4–I6: color palette, atmospheric conditions, water features, vegetation details, glacier appearance). The visual progression demonstrates convergence from an underspecified concept to a detailed, aligned output, illustrating how structured elicitation scaffolds preference articulation.

Original Prompt: Create the image of a sports car, it must have blue paint
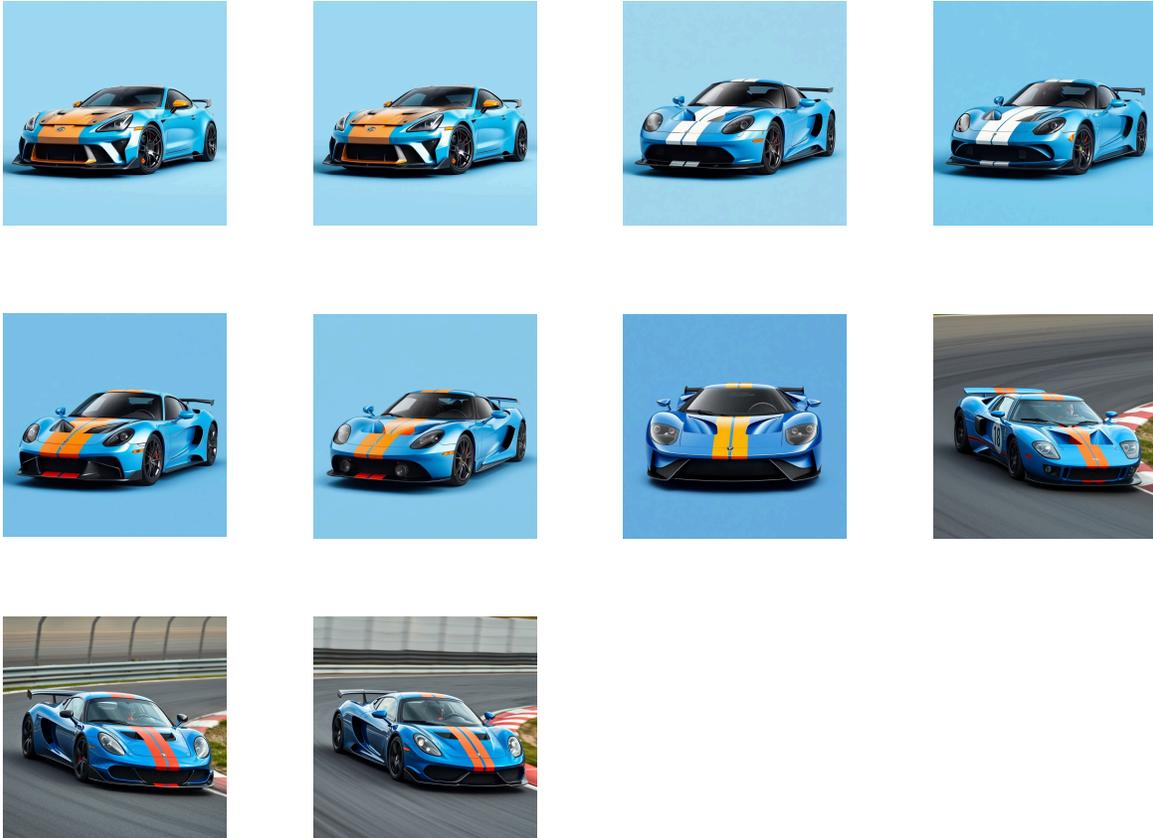
Final Prompt: Create the image of a European sports car on a curve of a circuit, it must be blue, with no stripes of other colors. realist style, with blur on the photo because of the speed

Figure 14: Examples of generated images from a Baseline participant. Across 10 generation attempts, the participant struggled to eliminate unwanted orange stripes despite progressively explicit constraints ("no stripes of other colors"). The final prompt grew from 12 to 35 words through iterative additions of negative specifications and contextual details ("European," "curve of a circuit," "blur because of the speed"), yet failed to approach the intended blue-only design. This illustrates the inefficiency of manual refinement: users must discover through trial which constraints matter, with no systematic guidance on resolving persistent misalignments.