Guided Creativity: AI Intermediation for Enhancing Originality and Quality in Visual Design

Xuekang Wu[†], Guy Aridor[†], and Artem Timoshenko[†]

[†] Kellogg School of Management, Northwestern University

March 2, 2026

Designers often improve the quality of their work by learning from successful exemplars. Yet this practice can trigger creative fixation, where exposure to existing solutions constrains originality in subsequent work. To resolve this tension, we introduce AI intermediation, a novel approach that leverages generative models to synthesize variations of leading designs. By presenting these variations instead of the original exemplars, our approach helps designers build on core design elements – visual motifs, color palettes, and structural composition – while maintaining variation in stylistic execution. We empirically validate this approach using a field experiment with professional designers in logo design contests. The results show that AI intermediation mitigates this tension: designers in this condition produce (1) higher-quality work than those with no exposure to exemplars, and (2) more original work than those directly exposed to original exemplars. Unpacking the mechanism, we find that AI variations effectively transmit core design elements, providing a foundation for the quality of the final designs. Furthermore, rather than serving as templates for imitation, the variations enable designers to translate underlying design logic into novel forms. Consequently, AI intermediation fosters creative collaboration, enhancing the human-to-human exchange of ideas.

Keywords: Generative AI, Creativity, Ideation, Crowdsourcing, Aesthetic Design.

1 Introduction

Good Artists Borrow, Great Artists Steal

Pablo Picasso

Learning from successful precedents is a fundamental approach to quality improvement and innovation across creative domains (Lidwell et al., 2010; Norman, 2013). Exposure to high-performing exemplars enables designers and organizations to discern effective strategies, align with evolving audience expectations, accelerate development by building on proven concepts, and raise baseline standards (Carpenter and Nakamoto, 1989; Cooper and Kleinschmidt, 1995; Von Hippel, 1986; Zhang et al., 2022). Many impactful innovations are not entirely novel, but rather creative reinterpretations or combinations of existing successful ideas. For organizations, Aerie built on the key message of Dove's Real Beauty campaign, yet forged a distinct brand identity that resonated well with its target audience (Maheshwari, 2016). Similarly, for individuals such as influencers, Ryan Trahan achieved massive success by studying MrBeast's viral content formats and adapting them with unique personal elements (Larner, 2022). These examples illustrate how inspiration drawn from success can be productively channeled into novel and successful creative outcomes.

However, learning from successful exemplars carries an inherent risk: creative fixation, where exposure to specific solutions causes creators to inadvertently anchor on the superficial features, diminishing the novelty and diversity of subsequent outputs (Berger and Heath, 2007; White and Argo, 2011). When creative fixation becomes widespread, the resulting homogenization can inflict significant damage. In the marketplace, this leads to visual saturation and audience fatigue, as initially distinctive concepts, such as Facebook's Corporate Memphis illustration style, become ubiquitous and lose their appeal (Huang, 2022). Strategically, visual convergence erodes the competitive advantage conferred by original designs, as pioneering brands see their unique identities diluted by look-alikes. This can lead to defensive legal measures, such as Oatly's lawsuits over packaging aesthetics and Apple's disputes

with Samsung regarding "slavish copying" of design (BBC News, 2021; Reuters, 2011). Beyond these competitive concerns, creative fixation confines designers to premature solutions instead of pursuing potentially superior alternatives in the broader design space, which limits brands' access to diverse stylistic options required for differentiation and customer appeal.

The challenge of balancing learning from successful exemplars against the risk of creative fixation often occurs in internal design processes within organizations, but it is particularly acute in crowdsourcing contests (Burnap et al., 2023; Terwiesch and Xu, 2008; Jiang et al., 2022; Mihm and Schlapp, 2019). Open contests, which allow participants to view leading submissions, typically yield higher average submission quality (Wooten and Ulrich, 2017a; Zhang et al., 2019). However, this transparency leads to significant imitation and a loss of originality due to fixation on early successful entries (Erat and Krishnan, 2012; Kornish and Ulrich, 2011; Hofstetter et al., 2021; Koh and Cheung, 2022). In contrast, blind contests, which withhold peer submissions, tend to foster greater originality and broader exploration but may result in submissions of lower average quality as learning opportunities are eliminated. This poses an important question: Can mechanisms be developed to facilitate learning from successful designs while simultaneously mitigating the detrimental effects of fixation to preserve creative exploration?

We propose AI intermediation, an approach that provides designers with *variations* of exemplars instead of the originals. The variations are created by a generative model that preserves core design elements, including visual motifs, color palettes, and structural composition, while ensuring variation in stylistic execution. By making the exemplars perceptually distinct, we help designers to internalize the underlying logic while discouraging the replication of surface features. This approach reframes Generative AI: rather than serving solely as a tool for individual augmentation, the generative model acts as a mediator, enhancing the effectiveness of information transmission between human designers.

We illustrate the design variations in Figure 1. The image on the left is the high-quality logo created by a professional designer ("original logo"), and we show four variations of

this logo on the right. All variations retain design elements (brand name, human figure and heartbeat line motifs, blue and orange colors, and flat design style) while differing in typography and specific rendering, providing visual distinctness. Designers can refine these variations or draw inspiration from them, though we show in our empirical application that they primarily do the latter, emphasizing the complementarity between the generative model and designers' creativity.

Figure 1: Illustrative Example of AI-Generated Variations

Notes: The original logo (left) for "Armor Health", featuring human figure, heartbeat line, blue, and orange palettes in a flat design, is transformed by our generative model into four distinct variations (right). These variations maintain core design elements (visual motifs, color palette, and structural composition) while differing in stylistic execution. We provide details about the generative model in Section 3.

Practical implementation of AI intermediation requires a specialized pipeline capable of generating high-quality variations. For our proof-of-concept application, we develop an approach that translates original designs into textual descriptions to capture core design elements, and then synthesizes new variations based on these descriptions. We fine-tune the image generation using a two-stage approach. First, the model learns foundational logo design principles through reconstructive training on professionally designed logos, ensuring the output resembles professional work. In the second stage, we align the model with high-performing design standards, calibrating it using survey-based quality measures.

The proposed generative pipeline creates visually coherent logo variations that are semantically similar to the original logos but visually distinct. We achieve these goals by

translating original logos into structured textual descriptions and then using these descriptions to guide image generation. Intuitively, our approach recognizes that language is an imperfect medium for communicating visual designs. Textual descriptions are sufficiently precise to capture the core design elements from the original logo (alignment), but cannot fully articulate all visual details (distinctiveness).

To empirically evaluate the AI intermediation approach, we conduct a field experiment within a real-world logo design contest. The experiment involves 292 professional designers recruited on a leading crowdsourcing platform. We randomize designers' access to logo exemplars in a creative brief across four treatment arms: Open, where designers view high-quality logos previously created for the focal brand; Blind, where designers view no logo exemplars; and two Variation conditions, where designers view AI-generated variations of original exemplars (hereafter referred to as 'variations').

We characterize the impact of the different contest types on the quality and originality of submissions. Quality is the primary objective of business creative processes. In the logo contest setting, the quality of the logos is often judged by the client who sponsors the contest, and for many brands, the primary purpose of the logo is to attract consumers via online ads. As such, we evaluate quality using survey-based ratings on how well logos attract clicks in online ads. Our second metric, originality, indicates whether designers are exploring novel ideas rather than converging on the provided exemplars. To evaluate originality, we calculate embedding-based and perception-based distance measures between submitted logos and exemplars from the brief. Our originality scores focus on high-quality designs, to separate originality from the quality dimension, and to mimic the practical setting where clients choose aesthetics after the initial quality screening.

Our findings confirm that AI intermediation successfully transmits valuable information, leading to quality improvements over the Blind condition, while spurring greater originality for high-quality submissions compared to the Open condition. The overall quality of designs from the Variation condition is on par with the Open condition, and both are about 13%

higher than the Blind condition. The originality of logo designs from the Variation condition is on the same level as the Blind condition and substantially higher than the Open condition. For brands, this translates into access to a richer pool of high-quality, diverse solutions, increasing the likelihood of identifying designs that not only meet objective quality criteria but also satisfy subjective aesthetic preferences.

We validate that the mechanisms underlying these results align with the conceptual framework of AI intermediation. First, submissions in both the Open and Variation conditions are more likely to incorporate the visual motifs, colors, and composition from the original exemplars – properties that we find are associated with higher quality. Second, we show that the set of exemplars in the Variation condition is less visually homogenized and that designers abstract from and build on these variations, rather than imitate them. Specifically, we show that (i) quality and originality do not differ when we show designers 1 or 4 variations per human exemplar and (ii) final submissions outperform professional refinements of the variations in both metrics. Overall, the evidence indicates that variations transmit useful information about the core design elements associated with high quality while introducing visual heterogeneity. This combination positions variations as creative springboards for designers, reducing visual fixation while preserving the information essential for guiding them towards high-quality designs.

While our proof-of-concept focuses on logo design, the conceptual idea of AI intermediation applies broadly to visual design settings where creativity is shaped by social learning. First, creative fixation is a pervasive challenge whenever creators build on each other's work. By deliberately injecting algorithmic variation, AI intermediation counters this tendency and helps sustain diversity in the design space. Second, our approach addresses growing concerns about AI-driven homogenization of creative outputs (Kleinberg and Raghavan, 2021; Castro et al., 2023; Anderson et al., 2024). Purely automated systems risk converging on stylistically similar solutions, undermining originality. In contrast, our framework preserves human agency: the generative model facilitates the exchange of ideas, rather than replacing human

creativity. Human judgment remains central in curating and building on the AI-variations, and our findings show that this interplay is crucial for performance. More generally, AI intermediation offers a scalable method for creative exploration in domains such as product aesthetics, advertising imagery, and web interface design, where human creativity and social learning are critical but could benefit from structured injections of variety.

The remainder of the paper proceeds as follows. Section 2 reviews the relevant literature and positions our contribution. Section 3 details the generative pipeline, including the model overview and the fine-tuning procedures. Section 4 describes the experimental design, hypotheses, and empirical findings from the field experiment. Section 5 presents a replication in the restaurant industry, confirming that AI intermediation remains effective even in domains where design conventions are less standardized and the solution space is less constrained. Finally, Section 6 concludes with managerial implications, limitations, and directions for future research.

2 Related Literature

The advancing capabilities of generative models have attracted growing research into their potential to augment human creativity. Studies have explored various modes of human-AI interaction, including generative models as an ideation partner, co-creator, or evaluation tool in settings such as story composition, advertisement creation, and artwork creation (De Freitas et al., 2025; Doshi and Hauser, 2024; Chen and Chan, 2024; Zhou and Lee, 2024). These studies find that iterative human-AI processes can often combine the strengths of humans and AI to outperform purely human or purely AI outcomes (Boussioux et al., 2024). These approaches generally focus on enhancing the creative output of an individual or a small, directly collaborating team.

These machine augmentation paradigms are being applied to solve business challenges. Scholars have used machine learning models to map brand attributes to visual logo characteristics for data-driven 'moodboarding', trained models for generating and screening auto-

motive aesthetic designs, and demonstrated how machine-driven shape morphing can yield more market-attractive forms (Dew et al., 2022; Burnap et al., 2023; Chen et al., 2023). Similar to the broader machine-augmented creativity literature, these applications typically involve machines directly assisting a human expert in their creative tasks or decision-making processes (Heitmann et al., 2024). A complementary stream of findings suggests caution: prompting individuals with machine-generated exemplars can improve the creativity of individual outputs, but may lead to homogeneity at the group level (Ashkinaze et al., 2025; Holzner et al., 2025; Meincke et al., 2025).

Our work introduces a novel paradigm using generative models not merely for individual augmentation or direct co-creation, but as an intermediary to foster collective creativity. Our proposed AI intermediation operates by abstracting and diffusing the core conceptual elements of human submissions to other humans, transforming the original ideas into visually distinct variations. This mechanism aims to facilitate indirect social learning by communicating successful concepts across participants without triggering direct fixation, which is often caused by exposure to peer work. Therefore, our proposed approach addresses a different set of challenges related to the collaborative creative processes.

The theoretical foundation for AI intermediation lies at the intersection of design cognition and social learning. Creative fixation, or design fixation, is a well-documented phenomenon where exposure to existing solutions limits the novelty of subsequent ideas (Jansson and Smith, 1991; Smith et al., 1993; Sio et al., 2015; Luo and Toubia, 2015). Mechanistically, exemplars bias the cognitive search process because they make specific surface features highly available in memory and thus disproportionately sampled during ideation. Prior research have attempted to mitigate this by reducing the fidelity or modality of the stimulus, such as using partial images or descriptions to minimize copyable details (Cheng et al., 2014; Vasconcelos and Crilly, 2016; Sarkar and Chakrabarti, 2011; Crilly, 2019). However, low-fidelity or non-visual representations often remove information useful for quality. This tension parallels the dynamics in Particle Swarm Optimization in optimization theory that

balances the exploitation of known successes with the maintenance of diversity required for effective co-search (Kennedy and Eberhart, 1995; Bratton and Kennedy, 2007). Inspired by this analogy, we propose a solution that preserves the structural fidelity of the exemplar (enabling exploitation) while varying the surface representation (sustaining co-search). Our approach achieves this by automatically generating variations that share the same core design elements as the original, thereby encouraging designers to generalize from patterns rather than mimic specifics stylistic execution, and offering a scalable alternative to human supervision.

This learning-creativity tension that motivates our specific empirical setting is well-documented within crowdsourcing contests. Open contests, where participants view competitors' submissions and feedback, facilitate observational learning, leading to higher average quality but also causing imitation and reduced originality (Wooten and Ulrich, 2017b; Hofstetter et al., 2021; Koh and Cheung, 2022). Conversely, blind contests isolate designers, fostering broader exploration and novelty but potentially limiting quality improvement due to the absence of learning signals (Erat and Krishnan, 2012; Kornish and Ulrich, 2011). These findings suggest the difficulty of simultaneously improving submission quality via learning and improving submission originality via designer creativity within the traditional crowdsourcing approaches.

3 Generative Model for Variations

Implementing AI intermediation requires a generative pipeline capable of high-fidelity synthesis. While powerful, current off-the-shelf models often struggle to adhere to established stylistic principles or interpret precise design requirements. The resulting visual artifacts can confuse designers rather than inspire (see Web Appendix A for an illustration). We thus developed a custom pipeline designed to satisfy three critical objectives.

First, the variations must achieve semantic fidelity, effectively transmitting the visual motifs, color palettes, and structural composition of the original exemplars to enable social

learning. Second, they must exhibit stylistic heterogeneity, ensuring visual distinctiveness to prevent creative fixation. Third, these variations must meet a baseline of perceptual quality, appearing as plausible, professional logos with minimal artifacts. This final constraint is essential for adoption; our focus groups with logo designers indicated a strong aversion to low-quality or overtly artificial outputs, which they deemed unlikely to provide meaningful inspiration.

To meet these criteria, we developed an integrated pipeline that decouples content from style. The pipeline first extracts core design elements by translating original logos into structured textual descriptions, isolating the semantic essence of a design from its specific visual rendering. These descriptions then serve as prompts for a fine-tuned text-to-image (T2I) model that synthesizes logo variations. Through a two-stage fine-tuning process, we optimize the model to produce outputs that are both structurally sound and aligned with high-performing design standards. We next detail the methodology employed in each stage.

3.1 Textual Description

The primary objective of the textual description stage is to accurately capture the original logo's core conceptual elements, and to format these concepts into structured prompts for the text-to-image generation. We construct structured logo descriptions using two complementary pieces: a brief summary that summarizes information from the creative brief and an open-form description of the original exemplar generated by image captioning models.

The brief summary explicitly represents key logo attributes and nonvisual meta-information derived directly from the creative brief. Specifically, this part of the prompt includes contextual details such as the brand name, industry, and high-level styles, combined with visually salient features such as colors, typography, and composition. To facilitate efficient model learning, we employ standardized 'trigger words' (e.g., 'logo_style', 'symbol_color', etc). These trigger words serve as indicators for logo features, guiding the model to establish systematic associations between textual descriptions and their corresponding visual outputs.

For example, we show a brief summary of a logo from Figure 1 below:

LogoAI, white background, brand_name "Armor Health", industry healthcare, logo_style flat, modern, symbol_color blue, orange, white, font_color blue

The open-form description complements structured prompts by capturing intricate visual-semantic details, including visual motifs, color paletts, and structural composition. To generate the nuanced narratives, we employ an off-the-shelf image captioning model, JoyCaption, which provides rich and holistic descriptions of visual arrangements and subtle stylistic nuances embedded within the logo (fpgaminer, 2025). Continuing the previous example, the description expands:

logo_object A digital logo for "ARMOR HEALTH." The text "ARMOR" is in large, blue uppercase letters, while "HEALTH" is in orange. A blue and orange emblem of a person in a dynamic pose. A heartbeat line graph is in orange.

The open-form textual descriptions control the information flow from original exemplars to variations: The more information descriptions contain, the more perceptually similar variations are to the original exemplar. The structured descriptions we use capture core design features of exemplars, yet are insufficient to articulate all visual details. We demonstrate this relationship in Web Appendix B.4.

We combine the brief summary and the description into a standardized textual prompt. The downstream T2I model is fine-tuned to generate logos using this prompt structure.

3.2 Generate Logos

To generate logo variations from structured textual descriptions, we fine-tune a pre-trained T2I model, FLUX.1-dev, using Low-Rank Adaptation (LoRA) (Black Forest Labs, 2024; Hu et al., 2022). This approach begins with a state-of-the-art base model and progressively adapts its capabilities to the specific context in two stages. The first stage focuses on instilling foundational design principles to generate visually reasonable logos. The second stage further improves the model's capabilities by optimizing for a specific dimension of output

quality (click attractiveness) using contrastive learning techniques. Both stages contribute to improving the model's interpretation of textual prompts.

We conducted focus groups with professional logo designers to study their established design process and the value of machine-generated solutions. The interviews highlighted two issues: First, machine-generated logos often fail to follow graphic design conventions, sometimes rendering elements too realistically or positioning them incorrectly. For example, in Figure 2A (left), the human figure overlays the letter "M", and there is insufficient contrast between the color of the human figure and the letter. On top of that, the color of "HEALTH" makes it barely visible in a white background. Second, these models can exhibit strong stylistic biases, such as the "clip art" tendency of Nano Banana Pro shown in Figure 2A (right), which produces outputs that appear generic and unprofessional (Google DeepMind, 2025). We provide additional examples of critical stylistic artifacts in Web Appendix A. These stylistic artifacts lower the logo quality, and are so frequent in current off-the-shelf models that designers cannot efficiently learn and iterate on ideas from the produced variations.

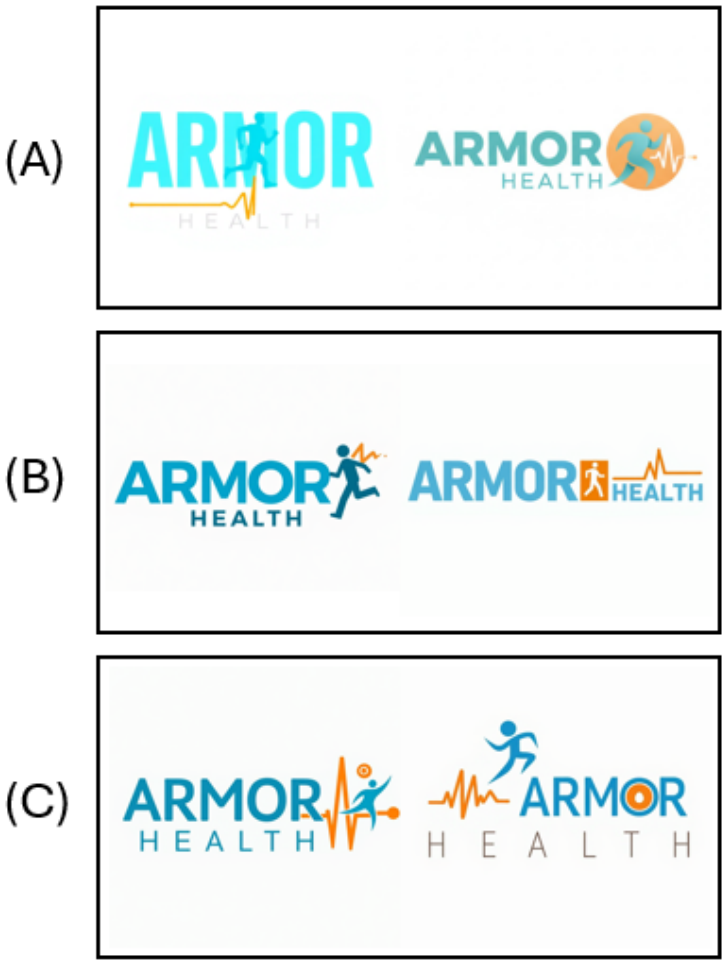
The fine-tuning is designed to address these challenges. The initial Logo LoRA (Section 3.2.1) directly addresses the challenge of instilling domain-specific knowledge and trains the model to generate visually reasonable logos that align with detailed textual prompts. The Logo LoRA helps ensure that the generated logos are not only relevant to the original concepts but also adhere to established design aesthetics. The subsequent Optimization LoRA (Section 3.2.2) further elevates the generative performance by learning from survey-based preference data on click attractiveness.

3.2.1 Logo LoRA

We first train the model to reproduce the graphic design conventions and various styles in logos. To do this, we curate a large dataset of professionally-designed logos and train the model to recreate these logos based on their textual descriptions.

Data. We acquired data from a crowdsourcing design platform to create a specialized

Figure 2: Logos Generated by Different Models



Notes: This figure shows variations generated by different models using the same prompt: (A) left logo is generated by FLUX.1-dev; Right logo is generated by Nano Banana Pro (2025); (B) outputs by pre-trained model + Logo LoRA; (C) outputs by pre-trained model + Logo LoRA + Optimization LoRA.

training set for this fine-tuning stage. Focusing on the restaurant industry as a proof-of-concept, we curated this dataset from past contest data, implementing several screening criteria to enhance training feasibility and mitigate the impact of low-quality images. At the contest level, we excluded contests requiring taglines or non-English brand names, given the known challenges of training accurate text generation with diffusion models. At the logo level, we removed images with low resolution and noisy backgrounds (e.g., logos on business cards). This screening process produced approximately 1,000 contests, from which we allocated 90% for the training set and 10% for hold-out validation; the training set

included about 25,000 logo images.

Fine-Tuning. We adopt a LoRA approach (Hu et al., 2022), a widely used technique for fine-tuning large models for specific applications. LoRA constrains the model training to a small subset of parameters, thereby retaining the original capabilities of the base model while adapting it to the specific task of logo generation. This approach is computationally efficient and ensures that the fine-tuned model can still remember concepts from the base model for logo generation. For example, if the fine-tuning logo dataset contains no examples with birds but the base model possesses prior knowledge of what a bird looks like, the fine-tuned model can still produce a visually coherent bird-themed logo.

During training, we finetune both the text encoder and the denoising network. The text encoder processes the structured textual prompt and converts it into an embedding that guides the image generation. We fine-tune the text encoder to help the model learn the trigger words introduced in Section 3.1. The denoising network is the primary image generator. It takes text embedding as a condition and learns to synthesize a logo that visually reflects the prompt. Web Appendix B.1 provides additional details about LoRA and latent diffusion models.

We illustrate the outputs from the Logo LoRA in Figure 2B. Compared to the outputs of off-the-shelf models, the outputs are more aligned with professionally designed logos from the crowdsourcing platform: they follow design principles better.

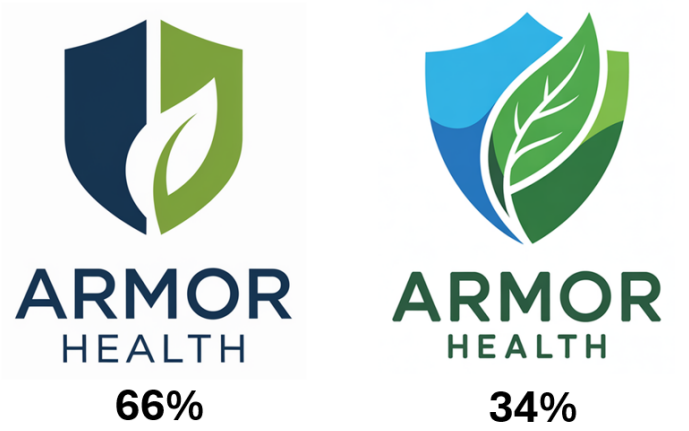
3.2.2 Optimization LoRA

In the Logo LoRA, we train the model to reconstruct professionally-designed logos. One challenge is that even within the curated set, there still exists variation in quality, and the model could be further improved if we train the model to yield more high-quality exemplars and avoid low-quality exemplars. In our proof-of-concept, we define the logo quality by how well it can attract clicks in display ads. Online advertising is a common use case for brand logos among small businesses. Our findings can be extended to other quality dimensions,

such as visual appeal, brand perceptions, or memorability.

Data. We measure click attractiveness using an online survey. We first select 50 pairs of logos from contests in our training data. These 50 pairs are selected so that they feature similar visual motifs (such as a fork and green leaves), but they are visually different designs. One illustrative example is shown in Figure 3. Each survey participant reviewed 25 logo pairs, and for each pair, indicated which logo they are more likely to click on in an online advertisement. We recruited 100 participants so that each pair receives 50 responses.

Figure 3: Illustrative Pairs for Optimization LoRA Training



Notes: This figure shows an example of the logo pairs used in the training of the Optimization LoRA. The two logos are similar in their core design elements, including motifs, color, and composition. However, the left logo has substantially higher click rates (66%) than the right logo (34%)

We assume that holding core logo design elements fixed, visual patterns not captured in the attributes can drive a logo to be marginally more or less click-attractive. In Web Appendix B.3, we compare the visual characteristics of the high-quality and low-quality logos (Liu et al., 2020; Zhang et al., 2022). The color brightness and symmetry of the logos are significantly different between the groups. This observation aligns with previous research that symmetric logos are perceived to be more preferable and that brightness shapes perceived organizational orientation that could interact with perception of restaurant brands, and this could contribute to higher click attractiveness (Luffarelli et al., 2019).

Fine-Tuning. To capture the visual patterns of logos with higher click attractiveness and

further align the image generation with logo design conventions, we use a separate LoRA with a contrastive loss (see details in Web Appendix B.1). The fine-tuning is constructed in a manner that for each pair of logos, the model learns to generate logos similar to the logo that performs better (higher click attractiveness) and different from the logo that performs worse. We illustrate variations generated by adding the Optimization LoRA to the Logo LoRA in Figure 2C.

3.3 Generative Pipeline Validation

We conducted extensive validation to confirm that the proposed generative pipeline yields logo variations that are semantically aligned with the original exemplars, visually distinct, and optimized for performance. First, we validate that the variations preserve the core design elements of the original exemplars, using both automated alignment metrics and human resemblance evaluations. Second, we demonstrate that the generated logos exhibit greater visual dispersion and less homogenization than the originals. Third, we show that the Optimization LoRA systematically increases color brightness and symmetry, characteristics that correlate with higher click attractiveness. Full details of these studies are provided in Web Appendix B.3.

4 Field Experiment

To empirically test whether AI intermediation can effectively facilitate learning across designers while avoiding creating fixation, we conduct a field experiment. The field experiment is implemented within a logo design contest, where we hired professional designers from freelancer.com. Specifically, we compare the performance of designers under AI intermediation to that of designers in two traditional types of contests: the Open condition with full exposure to exemplars and the Blind condition with no provided exemplars.

4.1 Study Design

The core of our experiment involves manipulating designers' access to a curated set of 60 logo exemplars for a small business healthcare brand. These logo exemplars were collected by the focal brand three years prior to our research, and received varying click attractiveness rates in our online survey, thus spanning the quality spectrum. We manipulated whether these exemplars are provided to professional designers as an inspiration to create a *new* logo for the same brand in a crowdsourcing design contest. The crowdsourcing design contest had a typical contest prize of $400 to incentivize participation by experienced designers.[1]

We illustrate the experimental conditions in Figure 4. At the beginning of the contest, participants were randomly assigned to receive access to a gallery, with each gallery corresponding to one experimental condition. All participants can view the creative brief that includes information about the healthcare brand and a textual description of client preferences (Appendix A provides an example). Specifically, each condition has the following gallery:

- Blind condition: Designers received only the contest brief, with no access to pre-seeded exemplars or their variations.
- Open condition: Designers viewed the contest brief alongside a gallery displaying the 60 pre-seeded exemplars and their quality rating.
- Variation(1) condition: Designers viewed the brief and a gallery presenting one AI variation (created by our model) for each of the 60 pre-seeded exemplars, alongside the original exemplars' ratings.
- Variation(4) condition: Designers viewed the brief and a gallery presenting four variations (created by our model) for each of the 60 pre-seeded exemplars, alongside the

[1]The focal brand obtained the initial logos in a private crowdsourcing design contest in 2019. These logos are not public to web search, thus designers in our study have no external access to logo exemplars unless provided by us. The focal brand was also not included in the training of the generative pipeline in Section 3.

original exemplars' ratings.

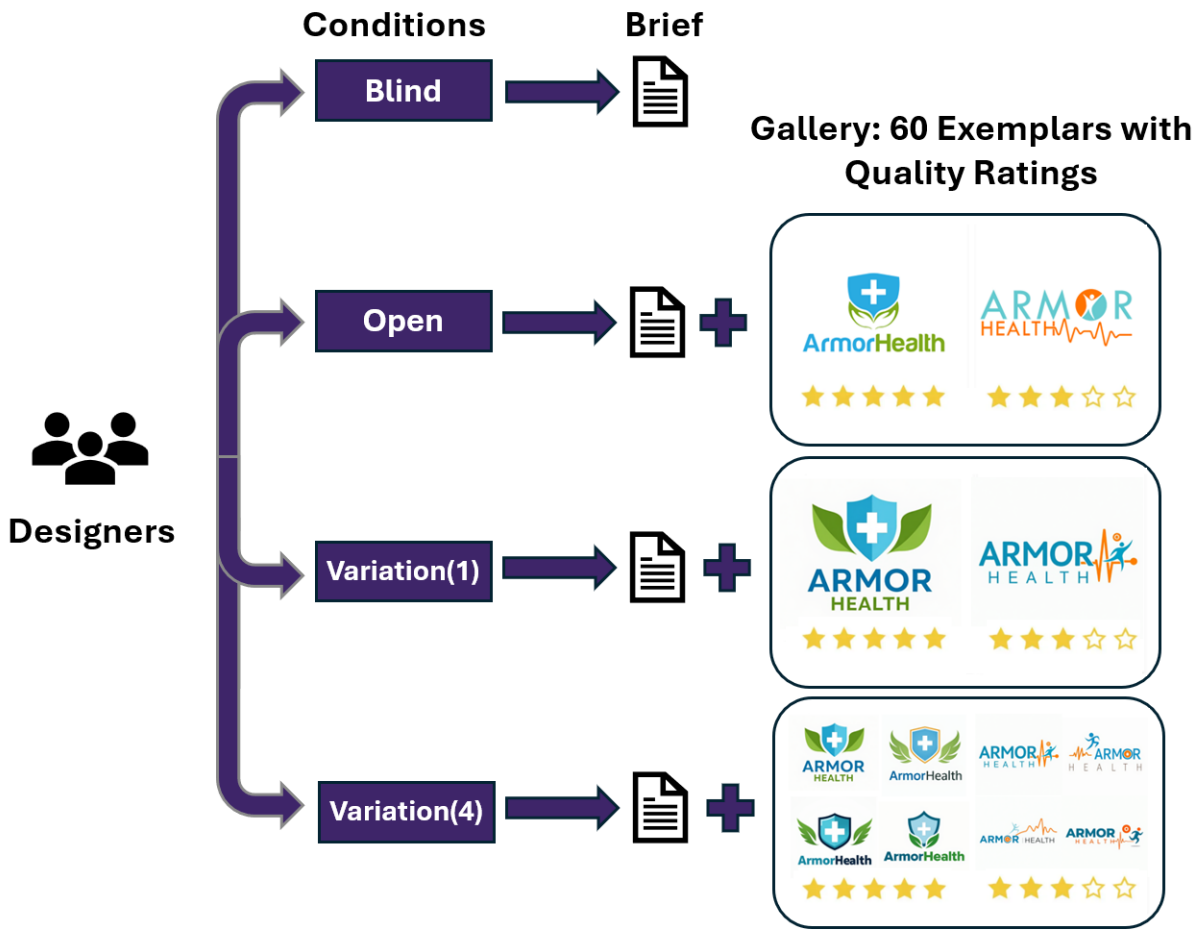
For the Open and the two Variation conditions, the exemplars (or their variations) are ranked by their quality ratings and presented on 5 pages, with 12 exemplars per page. Access to the galleries is restricted to assigned designers based on their designer IDs. Designers were informed that the goal was to create logos effective at attracting clicks in online display ads, and for the Open and Variation conditions, that the ratings of exemplars shown in the galleries reflected their click-attractiveness.

In this experiment, the Variation(1) condition is the focal treatment arm where AI Intermediation is introduced. The Open and Blind conditions serve as two control arms that correspond to the two traditional types of design contests. To understand the effectiveness of AI Intermediation, our following main analysis is on the comparison between the Variation(1) and two control conditions. We include the Variation(4) condition to test whether including more variations per exemplar can further improve contest outcomes compared to Variation(1). We present the comparison of these two conditions in Section 4.5. Hereafter, in this section, for simplicity, we call the Variation(1) condition the Variation condition.

The use of a gallery of logos within a brief balances the realism and methodological rigor. First, our study design follows the standard industry practice of including inspirational examples within a creative brief. Clients often provide examples of their favorite logos to indicate stylistic preferences. These examples can include internationally recognized brands such as BMW or Lacoste. Our study extends this idea by leveraging high-quality exemplars for the client's brand. Second, by providing a fixed set of exemplars from the outset, we ensure that every designer, regardless of when they join, operates within a consistent and controlled informational environment. This contrasts with traditional open contests, where designs are typically shown as they are submitted to the platform.

The contest ran for two weeks and followed typical specifications on the platform. To simulate realistic client feedback (ratings) during the contest, we randomly sampled 10 new submissions daily from each condition and provided ratings to designers. These ratings were

Figure 4: Experiment Design



Notes: Upon registration, designers were randomized into one of four conditions: (1) Blind, where they observed only the textual creative brief; (2) Open, where they observed the brief and 60 exemplars with quality metrics; (3) Variation(1), where they saw the brief and 60 variations (one variation per exemplar), along with the original exemplar's rating; or (4) Variation(4), which displayed three additional variations per exemplar beyond the Variation(1) stimuli.

sent through the platform in private messages, so that each designer could only see the rating for their own work if it was among the sampled submissions.[2] We ensured that designers remained unable to observe any information about other designers' submissions during the contest.

A total of 440 designers registered for the experiment, with 292 designers submitting at least one logo (Table 1). These participants were roughly equally distributed across the four conditions. While the Blind condition showed a slightly higher participation ratio among registered designers, this difference was not substantial compared to the Variation condition, suggesting that any unfamiliarity with the Variation condition did not significantly deter participation.

Table 1: Participants across Different Conditions

Condition	**Registered designers**	**Participating designers**	**Submissions**
Open	105	70 (66.7% of registered)	499
Variation(1)	105	71 (67.6% of registered)	575
Variation(4)	120	76 (63.3% of registered)	550
Blind	110	75 (68.2% of registered)	575

Notes: This table reports the sample sizes in different experimental conditions. Registered designers refer to those having viewed the contest information and creative brief. Registered designers are not required to participate in the contest. We define "participating designers" as designers submitting at least one design.

To characterize the participating designers and validate the randomization, we collected designer-level variables representing their experience and expertise from platform data. In Appendix B we provide the variable definitions and descriptive statistics. Participating designers had high client ratings and substantial experience, with an average of over 25 completed projects. Balance checks on the designer attributes across the conditions reveal no notable differences. Similarly, checks on participation patterns, including submission

[2]To collect these ratings, we measured click attractiveness using an online survey, similar to Section 3.2.2 and Web Appendix B.3. To ensure comparability, we benchmark the designers' submission to the original exemplars from the brief.

depth, entry timing, and continuous engagement, showed no substantial differences across conditions. This suggests the experiment was conducted with experienced designers under comparable conditions, allowing for a robust test of AI intermediation.

4.2 Experimental Hypotheses

The purpose of our experimental intervention is to test the effectiveness of AI intermediation at maintaining quality and mitigating creative fixation. We explicitly designed our generative pipeline in Section 3 to achieve the following two properties that we experimentally test:

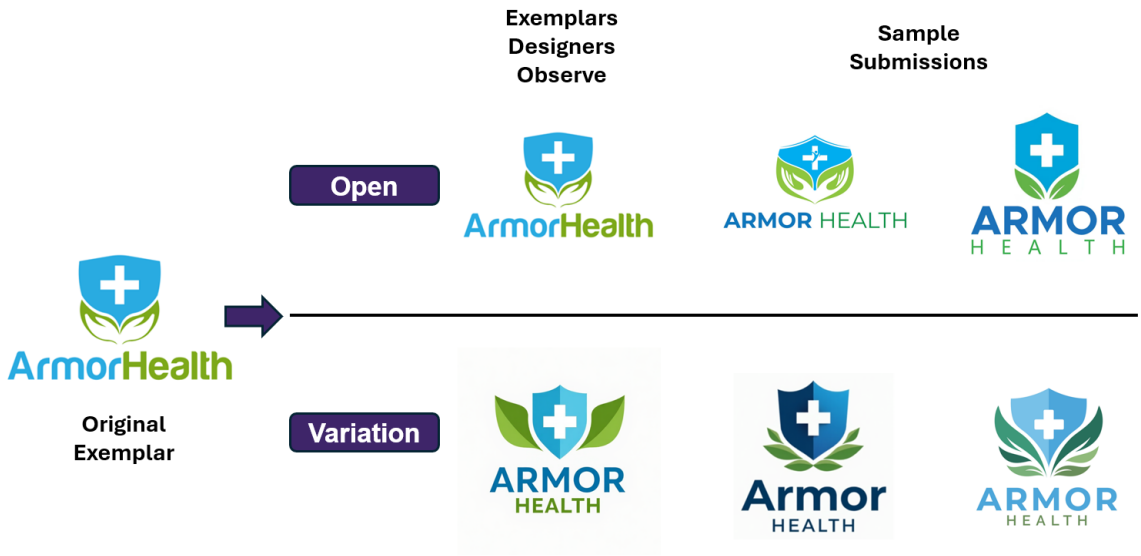
- Originality (Blind) $\geq$ Originality (Variation) $>$ Originality (Open)
- Quality (Blind) $<$ Quality (Variation) $\leq$ Quality (Open)

These predictions stem from a model of social learning where designers update their priors based on the gallery. Designers begin with prior beliefs about the core semantic elements and style of "successful" logos – from their previous experience and the creative brief – and use this to formulate their designs. In the Blind condition, designers rely on diffuse priors, resulting in the highest originality but lower average quality due to the lack of guidance. In the Open condition, the gallery provides a strong, homogenized signal for both semantic and visual features; this guides designers toward high-quality concepts but induces creative fixation (low originality).

AI intermediation is explicitly designed to decouple these two signals. As illustrated in Figure 5 and validated in Section 3.3, our generative pipeline successfully preserves core design elements, which are highly predictive of quality, while actively minimizing visual homogenization. Our field experiment tests the resulting behavioral impact on human designers: we expect that exposing designers to the Variation condition alters their beliefs about successful submissions. Specifically, we hypothesize that the AI-generated gallery transmits sufficient semantic value to retain the high quality seen in the Open condition,

while weakening the visual signal to mitigate creative fixation, thereby yielding significantly higher originality. We investigate these mechanisms in Section 4.5.

Figure 5: Diffusion of Leading Ideas in Open and Variation conditions



Notes: One the left, we show the original exemplar and AI-generated variations; all share similar motifs (shield, cross, leaves) and colors. The right column shows designer submissions. Submissions in the Open condition closely mimic the exemplar's form (such as leaves positioned at the bottom like holding hands). In contrast, submissions in the Variation condition abstract the core elements—blue/green palette, shield, cross, leaves—and recombine them in novel structural forms, illustrating the alleviation of fixation.

4.3 Originality

We first test the hypotheses regarding the incremental originality of submissions, defined as their distinctness from their most similar leading exemplars (hereafter referred to as "originality"). We measure originality using two complementary approaches: a scalable, embedding-based metric and a perception-based metric derived from human evaluations (Liu et al., 2020; Burnap et al., 2023; Compiani et al., 2025). Our findings are consistent across the two metrics; we report the embedding-based results in the main text and the perception-based results in Appendix C.

To construct the embedding-based metric, we use a pre-trained CLIP model to extract embeddings for all submissions and the 12 leading exemplars displayed on the first page of the

gallery (LAION, 2022).[3] CLIP captures both visual and conceptual information (Radford et al., 2021). The embedding-based originality of each submitted logo i is calculated as its minimum cosine distance to the 12 leading exemplars:

$$\text{Originality}_i = \min_{i' \in \text{Leading Exemplars}} 1 - \frac{e_i \cdot e_{i'}}{|e_i| * |e_{i'}|}$$

We then estimate the following regression analysis at the submission level, clustering standard errors by designer to account for multiple submissions from the same participant:

$$\text{Originality}_i = \sum_{c=1,2,3}^{C} \beta_c \mathbf{1}[\text{Cond}_{d(i)} = c] + \gamma \text{Day}_i + \delta^T X_{d(i)} + \epsilon_i \quad (1)$$

$$\epsilon_i = \eta_{d(i)} + \omega_i \quad (2)$$

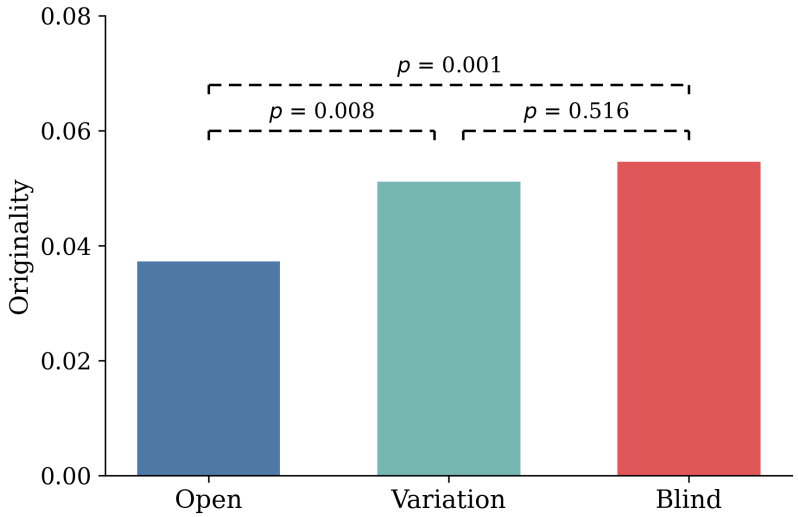
where Originality_i is the embedding-based originality of submission; β_c captures the treatment effects for the experimental conditions (Blind, Open, Variation); Day_i controls for the submission date; and $X_{d(i)}$ represents the designer's pre-experimental characteristics.

Figure 6 presents the results for the top-50 submissions with the highest quality ratings in each group. We focus on this high-quality subset because brands typically select from a pool of aesthetically appealing options. The Variation condition substantially outperforms the Open condition, achieving originality levels statistically indistinguishable from the Blind condition ($\Delta^{\text{Originality}}_{\text{Open, Variation}} = -0.014, p = 0.008$; $\Delta^{\text{Originality}}_{\text{Variation, Blind}} = -0.004, p = 0.516$). In Web Appendix C, we replicate the analysis using different definitions of the high-quality set and demonstrate that our main findings are robust.

These results confirm that AI intermediation successfully mitigates creative fixation, enabling designers to produce high-quality submissions that are significantly more original than those produced under direct exposure to leading exemplars. Notably, the similarity in originality between the Variation and Blind conditions suggests that our approach mitigates

[3]We focus on the first page because designers typically refer to top-rated logos to understand client preferences, and our gallery data confirms that most visits do not extend beyond the first page.

Figure 6: Mean Regression Estimates on Submission Originality across Conditions



Notes: This figure shows the point estimates from specification (1). P-values are for the contrasts between condition estimates.

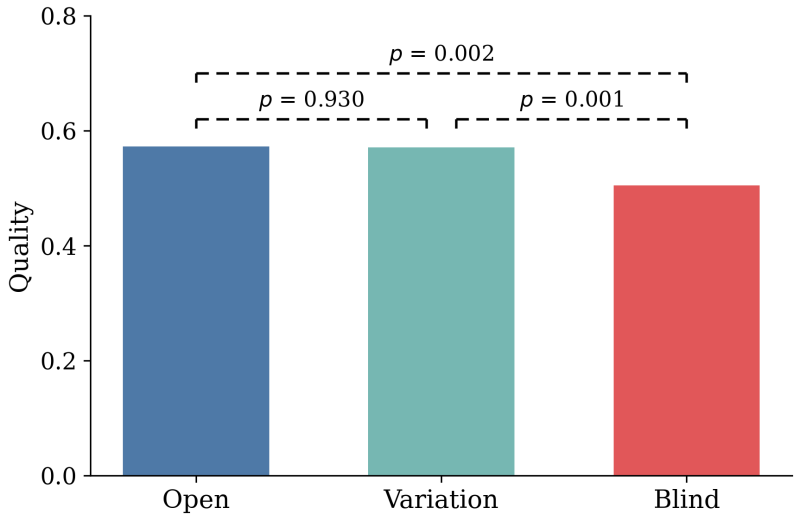
creative fixation as effectively as withholding exemplar information entirely.

4.4 Quality

We next evaluate the impact of experimental conditions on submission quality. To assess the average effects, estimate the same regression specification as in Equation (1), using click attractiveness (quality) as the dependent variable. We provide the full results in Web Appendix C. Pairwise contrasts in estimated coefficients in Figure 7 reveal that the Variation condition significantly outperforms the Blind condition, yielding a 13% increase in quality ($\Delta^{\text{Quality}}_{\text{Variation, Blind}} = 0.067, p = 0.001$), and is statistically indistinguishable from the Open condition ($\Delta^{\text{Quality}}_{\text{Open, Variation}} = 0.002, p = 0.930$). These results suggest that variations effectively transmit valuable information from leading designs, facilitating a level of learning and quality improvement comparable to direct exposure.

We further investigate whether AI intermediation improves the quality of top designs using quantile regression (details in Web Appendix C). Figure 8 presents the contrasts between conditions across the quality distribution. The x-axis represents the submission quality

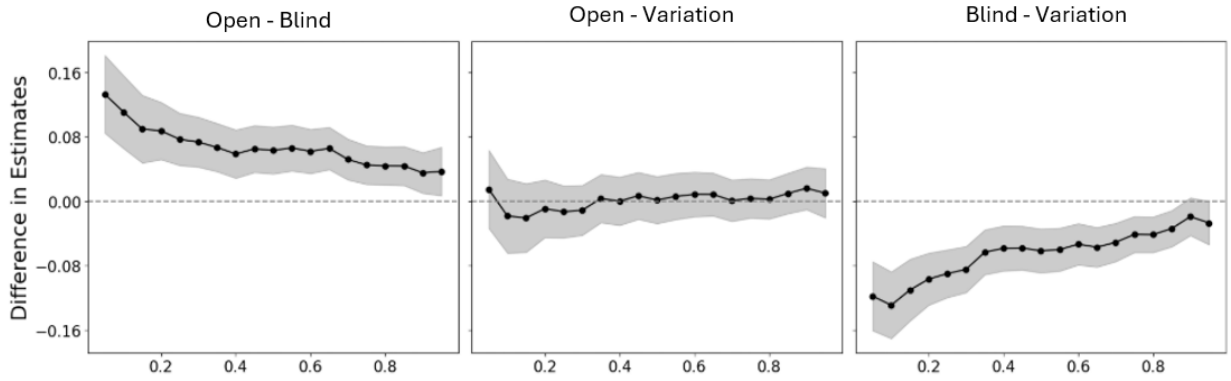
Figure 7: Mean Regression Estimates on Submission Quality across Conditions



Notes: This figure shows the point estimates from specification (1), with click attractiveness as the dependent variable. P-values are for the contrasts between condition estimates.

quantile, while the y-axis displays the difference in estimated coefficients. The Variation condition consistently outperforms the Blind condition, and remains statistically on par with the Open condition.

Figure 8: Quantile Regression Estimates on Submission Quality across Conditions



Notes: This figure shows contrasts of condition estimates with 95% CI derived from quantile regressions. The panels display differences in quality estimates for: Open vs. Blind (Left), Open vs. Variation (Center), and Variation vs. Blind (Right). The y-axis shows the difference in estimates, and the x-axis shows the quantile level.

In summary, AI intermediation not only enhances originality but also maintains high submission quality, achieving performance levels on par with full exposure while significantly

outperforming the Blind condition. This indicates that the learning benefits derived from observing leading exemplars are largely preserved in variations, demonstrating that increased creativity does not come at the expense of quality.

4.5 Mechanisms

We now explore the mechanisms behind the effectiveness of AI intermediation in maintaining the quality of the Open condition while achieving the originality of the Blind condition. Our results suggest a dual process where the generative pipeline successfully transmits high-value semantic information to solve the "blank page" problem, yet its visual presentation forces designers to engage in abstraction rather than imitation.

Regarding quality maintenance, a central premise of our design in Section 3 is that the galleries can facilitate the transmission of quality-related information. We find strong evidence that the visual motifs, color palettes, and composition from the original exemplars are indeed internalized by designers in the Variation condition. In Appendix E we show that specific core design elements – such as the medical cross, shield motif, and blue-green color palette – are highly predictive of quality in this context. We observe that these elements are equally likely to be present in the exemplars and variations in the Open and Variation conditions, respectively, and that designers in both conditions incorporated these high-value features at significantly higher rates than the Blind condition.[4] This pattern suggests that the AI variations successfully convey the "recipe" for high-quality designs. By transmitting the semantic constraints required for success without imposing the exact visual execution, the variations provide a starting point that allows designers to navigate the solution space effectively.

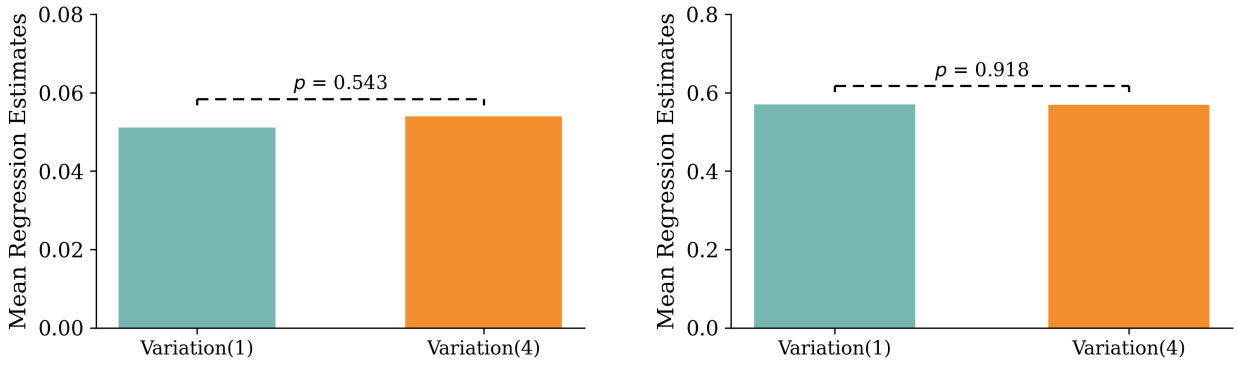
The divergence in originality between the Variation and Open conditions suggests a shift in cognitive processing. We posit an abstraction mechanism: the visual distance between the AI variations and the implied solution forces designers to extract the underlying semantic

[4] Additionally, we show that this same pattern holds for the aggregate alignment measures documented in Web Appendix B.3.

concepts and generate novel stylistic executions, rather than anchoring on specific surface details. This contrasts with an imitation mechanism, where designers might simply shift their fixation from the original exemplars to the AI variations, effectively displacing rather than mitigating the fixation.

To distinguish between these mechanisms, we conduct two complementary analyses. First, we test whether increasing the quantity of variations simply provides more distinct targets for imitation. If designers were primarily imitating the provided stimuli, providing four variations per exemplar should mechanically increase the diversity of the output compared to providing a single variation. However, as presented in Figure 9A, we find no significant difference in originality between the Variation(1) and Variation(4) conditions ($\Delta^{\text{Originality}}_{\text{Variation(1), Variation(4)}} = -0.003, p = 0.543$). This lack of sensitivity to the number of variations suggests that designers are not merely sampling from the surface features of the set, but are instead processing the underlying semantic signal which remains consistent across conditions.

Figure 9: Mean Regression Estimates on Submission Quality and Originality of Variation(1) and Variation(4) Conditions



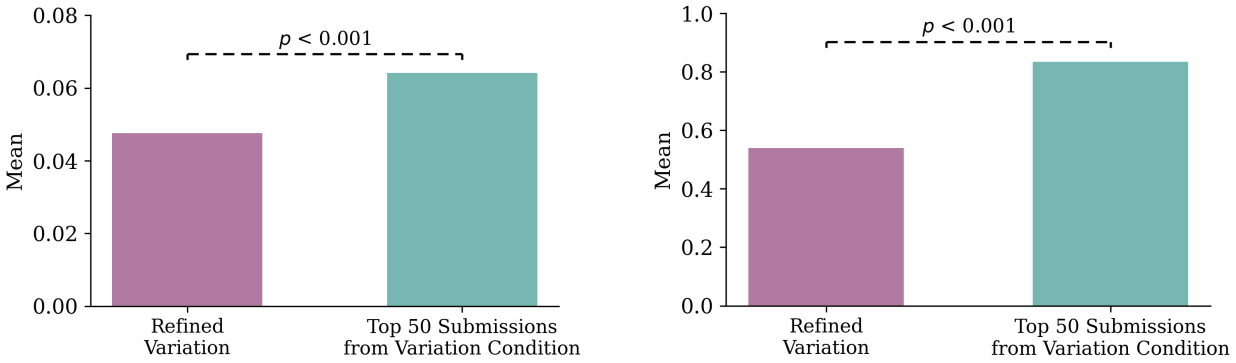
(A) Originality (B) Quality

Notes: This figure shows the point estimates from specification (1), with click attractiveness as dependent variable for panel (A) and originality as dependent variable for panel (B). P-values are for the contrasts between condition estimates.

Second, we investigate the extent to which human creativity adds value beyond the machine-generated output. If the imitation mechanism were the primary driver, we would

expect final submissions to closely resemble a polished version of the AI variations. To test this, we recruited professional designers to create "refined variations"—versions of the AI outputs polished to a ready-to-use state with minimal conceptual modification (see Appendix D for details). As shown in Figure 10, high-quality submissions from the Variation condition demonstrate substantially higher originality ($\Delta^{\text{Originality}}_{\text{Variation, Refined}} = 0.017, p < 0.001$) and quality ($\Delta^{\text{Quality}}_{\text{Variation, Refined}} = 0.295, p < 0.001$) than the refined variations. This substantial "human delta" indicates that designers use the variations as a creative springboard. They extract the high-value semantic parameters but traverse a search space that extends far beyond the visual suggestions of the AI.

Figure 10: Mean Regression Estimates on Quality and Originality of Refined Variations and High-Quality Submissions of Variation Condition



(A) Originality (B) Quality

Notes: This figure compares refined variations to the high-quality submissions in the Variation condition. The left panel shows regression estimates for click attractiveness. The right panel shows estimates for embedding-based originality. P-values are for the contrasts between condition estimates.

Taken together, these results support the abstraction mechanism. The effectiveness of AI intermediation stems from its ability to act as a soft constraint: it successfully transmits the semantic parameters required for quality, constraining the search to high-value regions, while its visual abstraction prevents the cognitive anchoring that leads to creative fixation.

5 Additional Application: Restaurant Logos

We conducted a replication study using a brand from the restaurant industry to evaluate the robustness of our main findings. This setting allows us to test AI intermediation in an industry with less standardized logo design conventions than healthcare. The replication study also altered the contest parameters, such as incentives and duration, to assess generalizability across a different distribution of participating designers. Recall that in Section 4, we have two Variation conditions, where designers observe one or four variations per exemplar. In this section, we keep the Variation(4) condition to help designers better abstract core design elements, and hereafter refer to it as the Variation condition.

Study Design. We closely followed the experimental design from Section 4. We conducted a logo design contest for a small business restaurant with three conditions: *Open* (full exposure to 60 rated exemplars of varying quality), *Blind* (no exemplar exposure), and *Variation* (exposure to four variations per exemplar). We followed the same designer recruitment process as in the main experiment.

We report the participation statistics in Table 2. A total of 485 designers registered for the contest, with 208 submitting at least one logo. Together, we gather 1027 submissions. Randomization checks, presented in Web Appendix D, confirm there are no significant differences in demographics or participation across conditions.

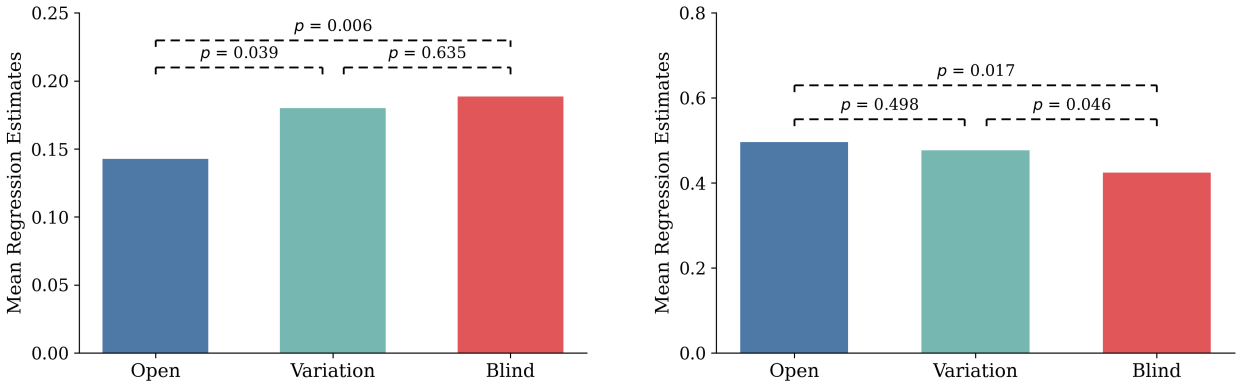
Table 2: Participants across Different Conditions

Condition	**Registered designers**	**Participating designers**	**Submissions**
Open	155	65 (41.9% of registered)	319
Variation	161	64 (39.8% of registered)	367
Blind	169	79 (46.7% of registered)	341

Notes: This table reports the sample sizes across experimental conditions. Registered designers refer to those who viewed the contest information and creative brief. Registered designers are not required to participate in the contest. We define "participating designers" as those submitting at least one design.

Results: Originality. We use the same embedding-based method to measure originality. Figure 11A presents the regression estimates for high-quality submissions (top in quality per condition) across conditions. The results are consistent with our previous findings: both the Variation and Blind conditions exhibit significantly higher originality than the Open condition. Detailed results and robustness checks using alternative definitions of high-quality submissions are provided in Web Appendix D.2.

Figure 11: Additional Application: Mean Regression Estimates of Submission Originality and Quality



(A) Originality (B) Quality

Notes: This figure shows the condition estimates under specification (1). Panel (A) shows estimates for originality; Panel (B) shows estimates for quality (click attractiveness). P-values are for the contrasts between condition estimates.

Results: Quality. Similar to our primary study, we use click attractiveness as the quality measure. Figure 11B shows the regression estimates on quality across conditions. The results mirror our earlier findings: two conditions providing exemplar information (Open and Variation) significantly outperforming the Blind condition, while showing no meaningful difference between themselves. Quantile regressions confirm this pattern holds across the entire quality distribution.

Discussion. These results demonstrate the robustness of AI intermediation's effectiveness across different industry contexts and design parameters. The success in both healthcare and restaurant contexts, despite the latter's less constrained visual norms, indicates that AI intermediation effectively facilitates creative learning and preserves quality even in settings

with flexible design conventions.

6 Conclusion

This paper proposes using generative AI as an intermediary between designers to facilitate creative learning: communicating the core design elements of successful concepts (motifs, colors, and composition) across designers without inducing creative fixation. We demonstrate the effectiveness of AI intermediation in a real-world logo design contest involving professional designers. Our proof-of-concept study highlights two primary effects. First, AI intermediation provides quality guidance: after observing variations of high-quality exemplars, professional designers produce higher-quality logos than those with no exemplar information. Second, AI intermediation helps to mitigate fixation: high-quality submissions from the Variation condition exhibit higher originality than those from the Open condition. Our analysis suggests that these effects arise because the variations (i) effectively transmit the core design elements associated with high quality and (ii) encourage designers to use the AI variations as creative springboards rather than directly imitating them.

These findings have important implications for visual design. By facilitating a portfolio of diverse high-quality concepts, AI intermediation can provide brands with more viable options that cater to different stylistic preferences and design objectives. We demonstrate this mechanism in a competitive design context, but the potential applications extend to collaborative environments. AI intermediation can act as a bridge for sharing creative information between design teams: When promising concepts are identified from market research or managerial guidance, variations can diffuse these concepts without causing creative fixation.

Limitations and Future Research

Future could explore more dimensions of 'variation'. In our proof-of-concept, we focus on brand logos and demonstrate the importance of capturing visual motifs, colors, and struc-

tural composition from the original exemplars. In more complex creative domains such as advertisements, product aesthetics, or architecture, variation can occur along multiple and orthogonal dimensions, such as shape, function, narrative tone, and cultural reference. An important line of inquiry, therefore, is whether generative models can achieve disentangled control over these specific attributes to better serve diverse design goals.

A second avenue for future work lies in expanding the AI intermediation paradigm to non-visual domains. Creative tasks such as product naming, slogan writing, and cross-modal branding also require complex mappings between abstract concepts and concrete representations. Investigating whether AI-mediated abstraction can foster creative learning in these domains could significantly broaden the scope and utility of the intermediation approach. Finally, the deployment of such systems requires a deeper understanding of human trust and cognitive reception of machine-generated content, particularly when AI is positioned not as a co-creator but as a neutral facilitator of creative communication.

References

Anderson, B. R., J. H. Shah, and M. Kreminski (2024). Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition*, pp. 413–425.

Ashkinaze, J., J. Mendelsohn, L. Qiwei, C. Budak, and E. Gilbert (2025). How AI ideas affect the creativity, diversity, and evolution of human ideas: Evidence from a large, dynamic experiment. In *Proceedings of the ACM Collective Intelligence Conference*, pp. 198–213.

BBC News (2021, August). Oatly loses trademark battle against Glebe Farm over oat milk.

Berger, J. and C. Heath (2007). Where consumers diverge from others: Identity signaling and product domains. *Journal of Consumer Research 34*(2), 121–134.

Black Forest Labs (2024, August). FLUX.1-dev (model card).

Boussioux, L., J. N. Lane, M. Zhang, V. Jacimovic, and K. R. Lakhani (2024). The crowdless future? Generative AI and creative problem-solving. *Organization Science 35*(5), 1589–1607.

Bratton, D. and J. Kennedy (2007). Defining a standard for particle swarm optimization. *2007 IEEE Swarm Intelligence Symposium*, 120–127.

Burnap, A., J. R. Hauser, and A. Timoshenko (2023). Product aesthetic design: A machine learning augmentation. *Marketing Science 42*(6), 1029–1056.

Carpenter, G. S. and K. Nakamoto (1989). Consumer preference formation and pioneering advantage. *Journal of Marketing Research 26*(3), 285–298.

Castro, F., J. Gao, and S. Martin (2023). Human-AI interactions and societal pitfalls. *arXiv preprint arXiv:2309.10448*.

Chen, B., J. Huang, M. Zhang, and L. Luo (2023). Does that car want to give me a ride? Bio-inspired automotive aesthetic design. *SSRN Electronic Journal*.

Chen, Z. and J. Chan (2024). Large language model in creative work: The role of collaboration modality and user expertise. *Management Science 70*(12), 9101–9117.

Cheng, P., R. Mugge, and J. Schoormans (2014). A new strategy to reduce design fixation: Presenting partial photographs to designers. *Design Studies 35*, 374–391.

Compiani, G., I. Morozov, and S. Seiler (2025). Demand estimation with text and image data. *arXiv abs/2503.20711*.

Cooper, R. G. and E. J. Kleinschmidt (1995). Benchmarking the firm's critical success factors in new product development. *Journal of Product Innovation Management: An International Publication of the Product Development & Management Association 12*(5), 374–391.

Crilly, N. P. (2019). Methodological diversity and theoretical integration: Research in design fixation as an example of fixation in research design? *Design Studies*.

De Freitas, J., G. Nave, and S. Puntoni (2025). Ideation with generative AI—in consumer research and beyond. *Journal of Consumer Research 52*(1), 18–31.

Dew, R., A. Ansari, and O. Toubia (2022). Letting logos speak: Leveraging multiview representation learning for data-driven branding and logo design. *Marketing Science 41*(2), 401–425.

Doshi, A. R. and O. P. Hauser (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances 10*(28), eadn5290.

Erat, S. and V. Krishnan (2012). Managing delegated search over design spaces. *Management*

Science 58(3), 606–623.

fpgaminer (2025, May). JoyCaption (GitHub repository).

Google DeepMind (2025). Nano Banana Pro: Gemini 3 Pro image model. Technical report, Google.

Heitmann, M., T. P. Jansen, M. Reisenbichler, and D. A. Schweidel (2024). EXPRESS: Picture perfect: Engaging customers with visual generative AI. *Journal of Marketing*.

Hofstetter, R., D. W. Dahl, S. Aryobsei, and A. Herrmann (2021). Constraining ideas: How seeing ideas of others harms creativity in open innovation. *Journal of Marketing Research 58*(1), 95–114.

Holzner, N., S. Maier, and S. Feuerriegel (2025). Generative AI and creativity: A systematic literature review and meta-analysis. *arXiv abs/2505.17241*.

Hu, E. J., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. (2022). LoRA: Low-rank adaptation of large language models. *The International Conference on Learning Representations (ICLR) 1*(2), 3.

Hu, M.-K. (1962). Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory 8*, 179–187.

Huang, L. (2022, March). Blue people and long limbs: How one illustration style took over the corporate world.

Jansson, D. G. and S. M. Smith (1991). Design fixation. *Design Studies 12*(1), 3–11.

Jiang, Z., Y. Huang, and D. R. Beil (2022). The role of feedback in dynamic crowdsourcing contests: A structural empirical analysis. *Management Science 68*(7), 4858–4877.

Johnson, J., R. Krishna, M. Stark, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei (2015). Image retrieval using scene graphs. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3668–3678.

Kennedy, J. and R. Eberhart (1995). Particle swarm optimization. In *Proceedings of ICNN'95-International Conference on Neural Networks*, Volume 4, pp. 1942–1948.

Kleinberg, J. and M. Raghavan (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences 118*(22), e2018340118.

Koh, T. K. and M. Y. Cheung (2022). Seeker exemplars and quantitative ideation outcomes in crowdsourcing contests. *Information Systems Research 33*(1), 265–284.

Kornish, L. J. and K. T. Ulrich (2011). Opportunity spaces in innovation: Empirical analysis of large samples of ideas. *Management Science 57*(1), 107–128.

LAION (2022). CLIP-ViT-bigG-14-laion2B-39B-b160k. Model trained by Mitchell Wortsman on stability.ai cluster.

Larner, H. (2022, July). Ryan Trahan's penny challenge buys a winning new YouTube format.

Lidwell, W., K. Holden, and J. Butler (2010). *Universal principles of design, revised and updated: 125 ways to enhance usability, influence perception, increase appeal, make better design decisions, and teach through design*. Rockport Pub.

Liu, L., D. Dzyabura, and N. Mizik (2020). Visual listening in: Extracting brand image portrayed on social media. *Marketing Science 39*(4), 669–686.

Luffarelli, J., A. Stamatogiannakis, and H. Yang (2019). The visual asymmetry effect: An interplay of logo design and brand personality on brand equity. *Journal of Marketing Research 56*(1), 89–103.

Luo, L. and O. Toubia (2015). Improving online idea generation platforms and customiz-

ing the task structure on the basis of consumers' domain-specific knowledge. *Journal of Marketing 79*, 100–114.

Maheshwari, S. (2016, March). Aerie's body positive message sent sales skyrocketing.

Meincke, L., G. Nave, and C. Terwiesch (2025). ChatGPT decreases idea diversity in brainstorming. *Nature Human Behaviour*, 1–3.

Mihm, J. and J. Schlapp (2019). Sourcing innovation: On feedback in contests. *Management Science 65*(2), 559–576.

Norman, D. (2013). *The design of everyday things: Revised and expanded edition.* Basic Books.

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PmLR.

Reuters (2011, April). Apple sues Samsung over 'slavish' copying of iPhone, iPad.

Sarkar, P. K. and A. Chakrabarti (2011). Assessing design creativity. *Design Studies 32*, 348–383.

Sio, U. N., K. Kotovsky, and J. Cagan (2015). Fixation or inspiration? A meta-analytic review of the role of examples on design processes. *Design Studies 39*, 70–99.

Smith, S. M., T. B. Ward, and J. Schumacher (1993). Constraining effects of examples in a creative generation task. *Memory & Cognition 21*, 837–845.

Terwiesch, C. and Y. Xu (2008). Innovation contests, open innovation, and multiagent problem solving. *Management Science 54*(9), 1529–1543.

Van der Maaten, L. and G. Hinton (2008). Visualizing data using t-sne. *Journal of Machine Learning Research 9*(11).

Vasconcelos, L. A. and N. P. Crilly (2016). Inspiration and fixation: Questions, methods, findings, and challenges. *Design Studies 42*, 1–32.

Von Hippel, E. (1986). Lead users: A source of novel product concepts. *Management Science 32*(7), 791–805.

White, K. and J. J. Argo (2011). When imitation doesn't flatter: The role of consumer distinctiveness in responses to mimicry. *Journal of Consumer Research 38*(4), 667–680.

Wooten, J. O. and K. T. Ulrich (2017a, Jan). Idea generation and the role of feedback: Evidence from field experiments with innovation tournaments. *Production and Operations Management 26*(1), 80–99.

Wooten, J. O. and K. T. Ulrich (2017b). The impact of visibility in innovation tournaments: Evidence from field experiments. *Available at SSRN 2214952*.

Zhang, S., D. Lee, P. V. Singh, and K. Srinivasan (2022). What makes a good image? Airbnb demand analytics leveraging interpretable image features. *Management Science 68*(8), 5644–5666.

Zhang, S., P. V. Singh, and A. Ghose (2019). A structural analysis of the role of superstars in crowdsourcing contests. *Information Systems Research 30*(1), 15–33.

Zhou, E. and D. Lee (2024). Generative artificial intelligence, human creativity, and art. *PNAS Nexus 3*(3), 1–52.

Appendix

A. Example of Creative Brief

Brand name: Armor Health

Armor Health is a healthcare company that provides comprehensive medical, dental, and mental health services around communities in the United States.

Keywords: Professional, Reliable

About the logo:
We are looking for a modern and sleek logo that is inviting yet eye-catching. We are open to ideas and any color scheme. Please include the text 'Armor Health' in the logo.
When submitting your designs, please **use a white background** and **do not use any mockup**.

IMPORTANT: We want to have a logo that can make our Facebook ads **more engaging and attract more clicks.**

To inspire you and help guide your designs, we provide ratings on logos that we previously collected in the gallery below. These ratings show how well logos attract clicks.

B. Designer Variables and Balance Checks

Table B.1 presents the summary statistics of designer-level variables of participating designers. Overall, designers exhibit common participation patterns as in other design contests: On average, designers submitted 7.5 designs, made submissions on 2.08 days, and entered the contest on Day 4, with entries concentrated at the beginning and end of the contest. We conducted pairwise t-tests on these participation measures and there is no statistically significant difference in any of these variables across the experimental conditions.

Table B.1: Designer Variables and Summary Statistics across Conditions

Variable	Description	Open	Blind	Var(1)	Var(4)
Reputation	System-generated past performance rating (0–5)	3.51 (2.36)	3.68 (2.11)	3.82 (2.04)	3.56 (2.21)
Professionalism	Avg. client rating of professional conduct (0–5)	3.53 (2.25)	3.69 (2.12)	3.83 (2.04)	3.57 (2.21)
HireAgain	Avg. client rating of rehire likelihood (0–5)	3.52 (2.24)	3.67 (2.11)	3.82 (2.04)	3.58 (2.23)
Quality	Avg. client rating of project quality (0–5)	3.53 (2.24)	3.69 (2.12)	3.83 (2.04)	3.58 (2.21)
NumJobs	Total number of completed projects	24.60 (44.37)	21.10 (46.70)	20.96 (44.04)	26.08 (56.54)
Reviews	Total number of client reviews received	23.51 (42.31)	20.62 (46.03)	20.45 (43.26)	25.26 (54.97)
HourlyRate	Designer-reported hourly rate	21.09 (18.01)	24.02 (31.50)	20.97 (20.53)	21.56 (18.36)

Notes: This table shows the designer-level demographics of participating designers across the four conditions. Standard deviations are in parentheses. Professionalism, HireAgain, and Quality are from client reviews when design projects are completed. Reputation is a weighted score calculated by the platform based on the three review dimensions. NumJobs refer to the number of completed design projects, and Reviews refer to the number of reviews received. HourlyRate is a designer self-reported hourly rate for design projects and it does not necessarily reflect the compensation designers receive.

C. Perception-Based Originality Measure

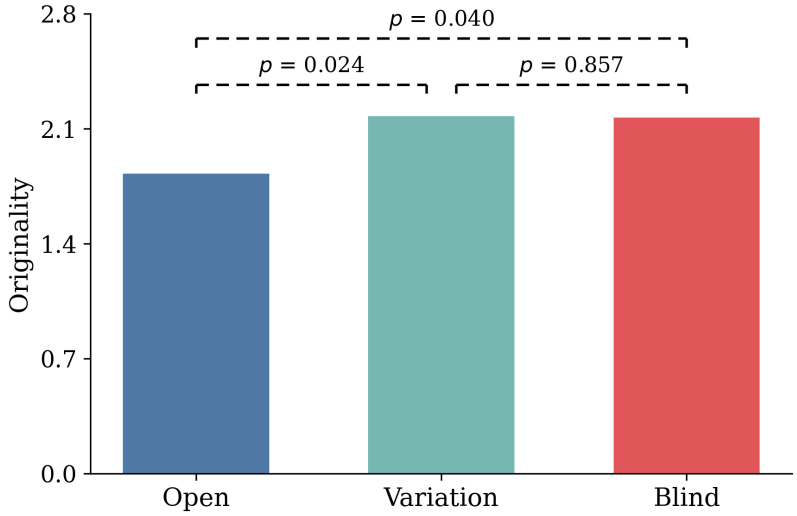
To complement the originality results in the field experiment, we additionally characterize the submission originality using a measure based on human perception. This measure captures survey-based evaluations of the visual designs rather than following the embedding approach.

Survey participants compared submissions on three dimensions: color palette, composition, and style. For each logo pair and dimension, 15 participants evaluated similarity on a 7-point Likert scale. To calibrate judgment, participants were shown two reference pairs at the beginning of the survey: one exhibiting high similarity and one exhibiting low similarity in the focal dimension of the survey. The core design elements in other dimensions are kept the same in these pairs.

The three dimensions capture partially distinct information: style is moderately correlated with color and composition ($\rho = 0.41$ and $\rho = 0.44$, respectively), while the correlation between the latter two is fairly low ($\rho = 0.09$). We use the mean over these three dimensions to define the perception-based originality score for each submission i:

$$\text{Perceived Originality}_i = \min_{i' \in \text{Leading Exemplars}} AVG(\Delta\text{Color}_{i,i'}, \Delta\text{Style}_{i,i'}, \Delta\text{Composition}_{i,i'})$$

Figure C.1: Mean Regression Estimates on Perception-Based Originality across Conditions



Notes: This figure shows the point estimates from specification (1), using the perception-based originality as the dependent variable. P-values are for the contrasts between condition estimates.

Figure C.1 presents the perception-based originality results. The Variation condition substantially outperforms the Open condition and is on par with the Blind condition:

$\Delta^{\text{Perceived Originality}}_{\text{Open, Variation}} = -0.351, p = 0.024$; $\Delta^{\text{Perceived Originality}}_{\text{Variation, Blind}} = 0.0257, p = 0.857$. These findings align perfectly with our conclusions for the embedding-based originality metric reported in the main text.

D. Logos in the Refinement Study

We recruited professional designers to refine the AI variations from the field experiment. The refinements focused on removing artifacts of the generative model, with minimal changes to the semantic and stylistic elements. We provide an illustrative example in Figure D.1.

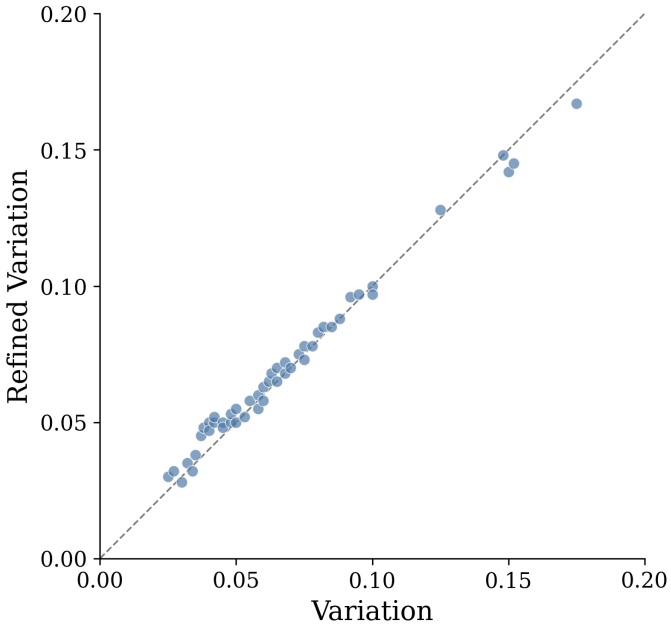
Figure D.1: Examples of Refined Variations

(a) Variations (b) Refined Variations

Notes: This figure shows two examples of refined variations. Panel (A) shows two machine-generated logos; Panel (B) shows the corresponding human-refined versions.

We validated that the refinement process introduced minimal conceptual modifications. Figure D.2 plots the embeddings-based originality scores of the AI variations against their refinements. The scores are highly correlated and cluster tightly along the 45-degree line. We subsequently use these refinements to investigate the underlying mechanisms: whether designers in the Variation condition substantially revised the provided exemplars or merely removed artifacts.

Figure D.2: Embedding-based Originality of Refined Variation vs. Variation



Notes: This figure compares the embeddings-based originality scores of variations and their refinements. Each point represents a pair of variation and its refined version. The value on the x-axis is the variation's originality score, and the value on the y-axis is the refined variation's originality score.

E. Core Design Elements and Quality Transmission

This section investigates the relationship between the core design elements (motifs, color, and composition) and logo quality. We establish three claims: (1) core design elements are meaningful predictors of logo quality, (2) the generative pipeline preserves these elements in the variations, and (3) designers in the Variation condition incorporate them at rates comparable to the Open condition (in the field experiment).

E.1 Association with Logo Quality

We first investigate whether logo design is predictive of the quality ratings. Our analysis focuses on three dimensions: motif, color, and composition. Motif captures visual symbols in a logo, such as a shield or medical cross. Color represents the primary colors a logo features (e.g., blue or black or green). We use ChatGPT 5.2 to extract the primary motifs

and colors for each design, restricting our analysis to elements present in at least 30 of the 2,199 total submissions. Composition refers to the spatial arrangement and geometry of the design. We use Hu moments for structural outlines and scene graphs for spatial relationships (Hu, 1962; Johnson et al., 2015). In total, our analysis incorporates 27 motifs, 10 colors, and 21 composition features. Table WA-B.1 in Web Appendix provides the complete list of variables.

Table E.1 evaluates the incremental contribution of these core design elements in explaining logo quality. The baseline model regresses submission quality on low-level computational aesthetic features alone. The baseline features include visual complexity, luminance and chromatic contrast, colorfulness, brightness, saturation, hue diversity, and horizontal symmetry. The computational aesthetic features account for 4.5% of the variation in quality ($R^2 = 0.045$). Incorporating motifs, colors, and composition individually increases the explanatory power by 8.6 to 20.5 percentage points. Adding all three dimensions jointly yields an R^2 of 0.360, demonstrating that the core elements can explain substantial variation in quality ratings.

Table E.1: Incremental Explanatory Power of Design Elements

Model	R^2	ΔR^2
Computational Aesthetic Features	0.045	—
+ Motif	0.250	0.205
+ Color	0.168	0.123
+ Composition	0.131	0.086
+ All three	0.360	0.317

Notes: This table reports the R^2 from OLS regressions of submission click rates on different sets of independent variables. The baseline model includes only low-level computer vision features. Each subsequent row adds one category of variables to the baseline. The final row includes all three categories jointly. ΔR^2 denotes the increase in R^2 relative to the baseline. All models are based on 2,199 observations.

E.2 Information Transmission

To identify which specific features matter, we apply Lasso regression including our 67 features and present the statistically significant predictors in a follow-up OLS regression in Table E.2, which highlights two important observations. First, we observe that the features most strongly correlated with quality, such as the shield motif and blue color, appear at high rates in exemplars. For instance, the shield motif appears in 30% of exemplars in the Open condition and 37% of variations. Similarly, the helmet motif is negatively correlated with quality, and it appears less frequently than other motifs in both original exemplars and variations.

Second, we observe that designers in the Open and Variation conditions internalize the quality information from exemplars and reflect it in their submissions. Namely, the shield motif appears in 64% and 58% of submissions from the Open and Variation conditions, respectively, which is substantially higher than the Blind condition where it only appears in 38% of submissions. Moreover, features that are common across healthcare logos but not differentially associated with quality or prevalent in the exemplars or variations in our context, such as the medical cross, show similar prevalence across all conditions. We conclude that variations transmit information about which specific features drive quality rather than generic industry conventions.

We can extend the feature-level analysis with the aggregate scores that combine information about motifs, colors, and composition. We proposed the alignment measures in Web Appendix B.2, as part of the generative pipeline validation. For each submission, we compute its alignment measures with the 12 original exemplars and then take the maximum as the submission's alignment score. Table E.3 reports mean alignment scores by condition at different quality thresholds. The submission alignment scores exhibit a significant association with logo quality ($\rho = 0.453, p < 0.001$). Across all submissions, both the Open (0.833) and Variation (0.823) conditions exhibit substantially higher alignment than the Blind con-

Table E.2: Quality-Relevant Design Elements: Predictive Power and Transmission

			Exemplar		Submissions		
Category	Index/Feature	Quality Cor.	Original	Variation	Open	Variation	Blind
Motif	Shield	0.339^{***}	0.300	0.367	0.635	0.581	0.381
	Medical cross	0.174^{**}	0.483	0.433	0.577	0.539	0.518
	Helmet	-0.080^{***}	0.083	0.067	0.014	0.033	0.016
Color	Blue	0.258^{***}	0.950	0.950	0.930	0.901	0.892
	Red	-0.154^{***}	0.200	0.233	0.128	0.129	0.115
	Yellow	-0.142^{***}	0.000	0.017	0.024	0.064	0.045
	Black	-0.110^{***}	0.200	0.183	0.214	0.151	0.228
Composition	Vertical Layout	0.182^{***}	0.950	0.963	0.942	0.964	0.873
	N. Components	0.170^{***}	12.4	12.3	11.8	11.9	10.7

Notes: The upper panel shows alignment indices; The lower panel shows binary features for motifs and colors. Quality Cor. reports pairwise Pearson correlations with submission quality ($N = 2199$). Exemplar columns show feature prevalence among the 60 exemplars and their AI-generated variations. Submission columns show feature prevalence or mean index values in designer submissions by experimental condition. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

dition (0.776). The Open and Variation conditions show comparable alignment, confirming that variations successfully transmit core design elements from original exemplars to final submissions.

A more revealing pattern emerges when we condition on submission quality. As we restrict to higher-quality subsets, the alignment gap between non-Blind and Blind conditions narrows: the Variation–Blind difference shrinks from 0.047 (all submissions) to 0.027 (top 200) to just 0.012 (top 50), and the Open–Blind gap follows a similar trajectory (0.057 to 0.030). The highest-quality Blind designers converge on similar design elements despite receiving no exemplar information. This suggests that variations guide a broader share of designers toward quality-relevant design elements.

Table E.3: Submission Mean Alignment Scores by Condition

	Alignment Score			Δ vs. Blind	
Subset	Open	Variation	Blind	Open	Variation
All Submissions	0.833 (0.085)	0.823 (0.077)	0.776 (0.110)	0.057	0.047
Top 200	0.856 (0.077)	0.852 (0.055)	0.825 (0.088)	0.031	0.027
Top 100	0.866 (0.076)	0.852 (0.054)	0.829 (0.089)	0.037	0.023
Top 50	0.872 (0.074)	0.854 (0.056)	0.842 (0.094)	0.030	0.012

Notes: Alignment scores measure the aggregate similarity between submissions and leading exemplars across motif, color, and composition (see Table WA-B.2 for construction). Standard deviations in parentheses. Top X refers to the X highest-quality submissions per condition ranked by click attractiveness. Δ vs. Blind reports the difference in mean alignment scores relative to the Blind condition.

Web Appendix to

Guided Creativity: AI Intermediation for Enhancing Originality and Quality in Visual Design

March 2, 2026

Web Appendix A. Logos Generated by Off-the-Shelf Models

In this section, we provide additional justification for using our custom generative model over existing off-the-shelf models. We begin with Figure WA-A.1, which shows variations produced by FLUX.1-dev. It is the natural benchmark as one of the most advanced open-source image-to-image models and the base for our pipeline. In this example, the four variations are similar in form, failing to meet our requirement that variations should be visually distinctive from the original logo.

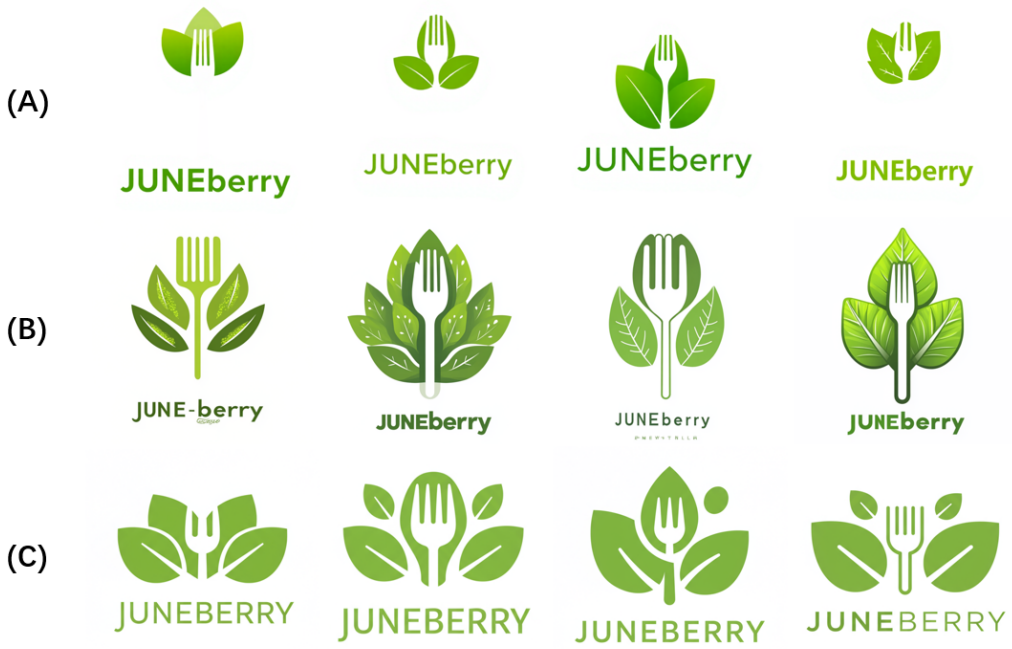
To formally assess whether our model achieves greater visual distinctiveness than image-to-image generation, we conducted a validation study using 60 randomly sampled logos from different brands. For each logo, we generated four variations using both our model and FLUX.1-dev, and measured distinctiveness as the average CLIP embedding distance between each set of variations and the original logo. A paired t-test confirms that our model produces substantially more distinctive variations ($\Delta Mean = 0.0414$, SE $= 0.0008$, t $=$ 51.102, $p < 0.001$).

Beyond visual distinctiveness, off-the-shelf text-to-image models exhibit additional shortcomings when prompted with our structured textual descriptions (Figure WA-A.2). FLUX.1-dev (Panel A) violates basic design principles: the fork lacks color contrast against the background, and the spacing between graphic elements and typography is excessive. Midjourney (Panel B) produces designs that are too detailed and complex to function as logos. Imagen (Panel C) generates outputs in a uniform style that diverges from the original logo.

Figure WA-A.1: Variations Generated by Image-to-Image Model

Notes: This figure shows the variations generated by the image-to-image approach. The original logo is on the left, and the four images on the right are variations generated using FLUX.1-dev.

Figure WA-A.2: Variations Generated by Off-the-Shelf Text-to-Image Model



Notes: This figure shows variations generated by different text-to-image models. The variations are all generated using the structured description of the original logo in Figure WA-A.1. (A) shows the outputs of FLUX.1-dev. (B) shows the outputs of Midjourney. (C) shows the outputs of Imagen.

Web Appendix B. Generative Model Development and Outputs

In this section, we provide details of the development of the generative pipeline and a closer look at the variations generated by the pipeline. We begin by providing details of the model architecture of the text-to-image model, and show how the length of descriptions generated by the image-to-text model can influence the variations; we then provide formal validation of the variations; we conclude by proposing measures of the alignment of the variations to the exemplars and use them to show that alignment is achieved.

B.1 Technical Details in Model Training

The pre-trained model in our generative pipeline is a diffusion model. Diffusion models work by reversing a diffusion process to synthesize data. The model training process is shown in Figure WA-B.1.

The image is initially encoded to an image latent, and then a forward diffusion process gradually adds noise to the latent, transforming it from the initial state z_0 to a Gaussian noise z_T. At time step t, the noised latent is:

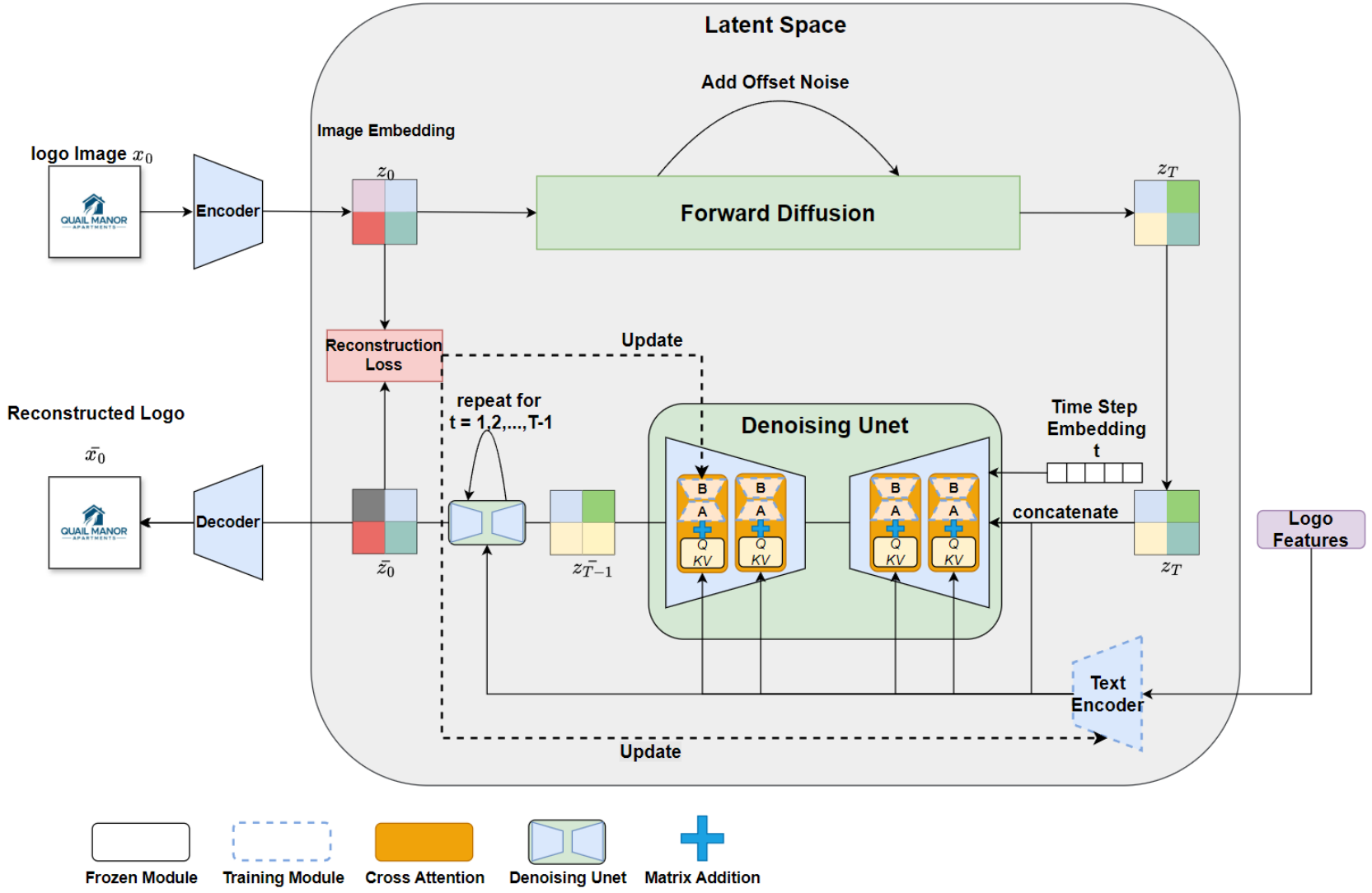
$$z_t = \sqrt{1-\alpha_t} z_0 + \sqrt{\alpha_t}\epsilon$$

Where ϵ is a Gaussian noise. The goal of Diffusion models is to learn to denoise the added noises so that a noisy state z_T can be reversed back to an image latent z_0. Therefore, at t, the loss is

$$||\epsilon - \epsilon_\theta(z_t, c, t)||^2$$

Here, θ is the model, c is the condition (i.e., embedding of the prompt). The denoised initial state z_0 is then decoded to obtain the final image. The loss is, in essence, a reconstruction loss of the original image. While minimizing the loss in the training, the model is learning to

Figure WA-B.1: Illustration of A Latent Diffusion Model



Notes: This figure shows the training process of a canonical latent diffusion model.

reconstruct the original logo as closely as possible given the logo description (prompt), thus implicitly forcing the model to learn graphic design principles and to align with the prompt.

For fine-tuning, we use LoRA, a method that allows efficient adaptation of pre-trained models for downstream tasks (Hu et al., 2022). Suppose the cross-attention layer of the pre-trained model is $W_0 \in \mathbb{R}^{d\times k}$, where d, k are the original and output dimensions respectively. LoRA trains ΔW to minimize the denoising loss. It is efficient as it decomposes ΔW as $\Delta W = BA$, where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$, with r being much smaller than both d and k. Intuitively, LoRA is compressing the updated information to low-rank matrices. During inference, the weights of the new model $\theta + \theta_{\text{Lora}}$ are $W_0 + \Delta W$.

In training of Logo LoRA as described in Section 3.2.1, we set the learning rate of the denoising network to be 6×10^{-6} and text encoder network to be 3×10^{-6}; r to be 128; batch size to be 10; and a cosine learning rate scheduler. The training converges at around 12 epochs, and takes less than 2 days on an A100 GPU.

For Optimization LoRA, instead of reconstruction loss, we propose a contrastive denoising loss:

$$\gamma(||\epsilon_p - \epsilon_{\theta+\theta_{\text{Click}}}(z_t, c, t)||^2 + ||\epsilon_n - \epsilon_{\theta-\theta_{\text{Click}}}(z_t, c, t)||^2)$$

Here, each loss is calculated as the sum of the loss for the noise added to the positive sample ϵ_p and the noise added to the negative sample ϵ_n. γ represents the level of click rate differences between the two logos and is calculated as $ln(\frac{a}{b})$, where a is the larger click rate in a pair, and b is the smaller rate. c represents the common prompts of the two logos. The gradients of θ_{Click} with respect to the contrastive loss are weighted by γ's of the pairs, meaning that the model learns more information from the pairs with larger click rate differences.

In training of Optimization LoRA as described in Section 3.2.2, we only update the denoising network, and the learning rate is 1×10^{-5}. The dimension of the Optimization LoRA is set to be 16. We use a cosine learning rate scheduler and batch size of 1. We train

the model for 20 epochs. During inference, we set the weight of Optimization LoRA to be 1.

B.2 Alignment Between Original Logos and Variations

We measure alignment between original logos and AI-variations using two approaches. The first approach uses CLIP embeddings to capture the logo's visual form, and then calculates distances between the logos in the embedding space. We use this approach to evaluate visual homogeneity of the generated logos in model validation.

The second approach decomposes the logo design along the three dimensions: motif, color, and composition. For each dimension, we extract relevant variables for each logo and characterize alignment between logo designs by calculating the overlap. We also average their standardized values to create composite indices. We summarize dimensions and their operationalization in Table WA-B.1, and the alignment measures in Table WA-B.2. These alignment measures help us validate that the generative pipeline preserves the core design elements to facilitate information exchange in AI intermediation.

B.3 Generative Model Validation Details

We provide validation that our pipeline produces variations that meet the objectives listed in Section 3. Study 1 validates that the variations generated by our generative pipeline are aligned to exemplars in core design elements, and thus achieving semantic fidelity. Study 2 shows that the variations are more visually heterogeneous than exemplars; Study 3 compares the performance of two fine-tuning steps and demonstrates that the Optimization LoRA enhances the quality of the generated variations.

Study 1. We apply the alignment measures developed in Appendix B.2 to validate that our generative pipeline produces variations that preserve the core design elements of the original exemplars. We compute alignment scores for 240 variations of 60 exemplars from the Armor Health experiment. Figure WA-B.2 provides illustrative examples at high, medium, and low alignment levels. In the high-alignment case, the variation preserves not only salient motifs

Table WA-B.1: Overview of Variables for Core Design Elements

Dimension	**Variables**
Motif	brain, caduceus, circle, dna helix, family, hand, heart, heartbeat line, helmet, house, human figure, leaf, letter a, letter h, medical cross, shield, snake, sparkle, staff, star, stethoscope, swoosh, sword, tooth, tree, triangle, wings
Color	black, blue, gray, green, lime, orange, purple, red, white, yellow
Composition	Hu moments: Seven rotation-, scale-, and translation-invariant image moments Scene graph features: Number of significant components; total spatial relationships; vertical layout present; horizontal layout present; containment present; overlap present; largest component dominance; component area entropy; horizontal spread; vertical spread; mean pairwise distance; above/below count; left/right count; containment count

(shield, cross, leaves) and the overall color palette, but also subtle spatial relationships such as leaves flanking the shield. In the medium case, salient motifs are retained but arranged differently. In the low case, the color palette and overall composition are preserved, but a motif that a heart forms the body of a human figure is lost entirely.

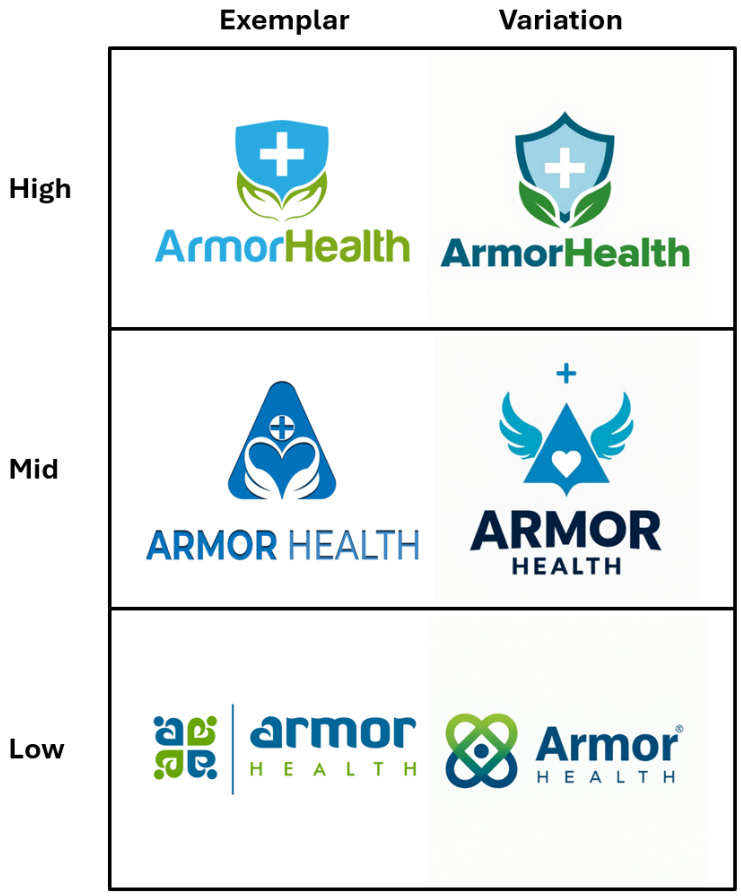
To formally benchmark alignment, we compare exemplar-variation (E-V) scores to an exemplar-exemplar (E-E) baseline computed across all pairwise combinations of the 60 exemplars. These E-E pairs represent logos designed for the same brand but without structural relationship to one another, and thus provide a natural benchmark. Table WA-B.3 presents the results. Across all measures and indices, E-V alignment substantially exceeds the E-E baseline, confirming that the generative pipeline produces variations that are meaningfully more aligned with their source exemplars than unrelated logos from the same brand.

We additionally validated alignment using survey-based measures with human partic-

Table WA-B.2: Overview of Alignment Measures

Dimension	**Measure**	**Description**
Motif	Dice coefficient	Overlap of distinct extracted visual symbols (e.g., "cat", "shield")
	Cosine similarity of embeddings	Conceptual proximity calculated via cosine similarity between text embeddings of extracted motifs
Color	Dice coefficient	Overlap of primary color sets
	Color moment distance	Euclidean distance between color distributions (mean, standard deviation, skewness) in CIELAB space
	Atmospheric style index	Aggregated differences in high-level aesthetic features: brightness, contrast, vibrance, and warmth
Composition	Scene graph edit distance	Operations needed to align spatial relationship graphs (e.g., "Icon LeftOf Text")
	Hu moments distance	Similarity of global geometric shape based on invariant statistical descriptors of the design's binary silhouette

Figure WA-B.2: Different Levels of Alignment in Core Design Elements



Notes: This figure presents three pairs of exemplars and their corresponding variations. From top to bottom, they have high, mid, and low scores in the overall alignment measure (avg. dimension indices).

.

Table WA-B.3: Comparison of Alignment Measures between E-V and E-E Pairs

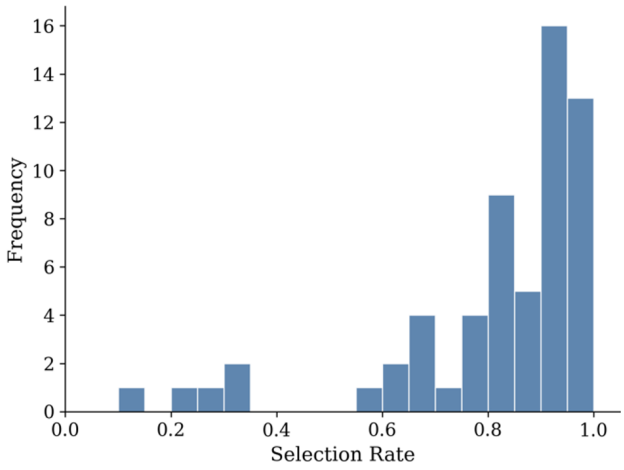
Measure	E-E Mean (SD)	E-V Mean (SD)	t	p
Individual Measures				
Motif Dice	0.163 (0.220)	0.610 (0.288)	28.37	< 0.001
Motif Embedding	0.940 (0.025)	0.973 (0.026)	18.62	< 0.001
Color Dice	0.549 (0.226)	0.854 (0.183)	20.03	< 0.001
Color Moments	0.936 (0.040)	0.953 (0.032)	6.56	< 0.001
Atmospheric Style	0.923 (0.082)	0.974 (0.027)	9.65	< 0.001
Scene Graph	0.721 (0.199)	0.773 (0.172)	3.79	< 0.001
Hamming Similarity	0.701 (0.050)	0.769 (0.041)	17.15	< 0.001
Composite Indices				
Motif Index	0.420 (0.162)	0.729 (0.205)	26.71	< 0.001
Color Index	0.701 (0.156)	0.871 (0.089)	16.47	< 0.001
Composition Index	0.559 (0.182)	0.698 (0.145)	11.30	< 0.001
Overall Indices				
Avg. Indices	0.560 (0.100)	0.766 (0.093)	30.15	< 0.001

Notes: This table shows the scores on proposed alignment measures of the E-E benchmark and E-V alignment groups. The upper panel shows the scores that are used to construct the dimension indices; the middle panel shows the dimension indices; and the lower panel shows the four overall alignment scores.

ipants. Specifically, we collected the perceived resemblance of variations to the original exemplar against that of the most similar logos from the same design contest. We sampled 60 original exemplars from different brands. For each, we used our pipeline to generate one variation. We also identify a retrieval-based benchmark: among logos submitted by other designers in the same contest, we select the one with the smallest CLIP cosine distance to the source logo. We excluded the source designer, who may submit near-identical designs multiple times. Survey participants were then presented with pairs consisting of our variation and the retrieval-based most similar logo, and asked to select which of the two better resembled the original logo. Each pair was evaluated approximately 100 times.

Figure WA-B.3 demonstrates strong alignment between the original exemplars and their variations. The mean selection rate for variations was 0.812 (t=52, significantly different from 0.5), indicating that subjects perceived variations as sharing substantially more resemblance to original exemplars than the most visually similar alternative from the same contest. This supports the conclusion that our pipeline effectively captures and re-generates the core design elements of the original exemplars.

Figure WA-B.3: Resemblance to Exemplars of Variations vs. Most Similar Logos



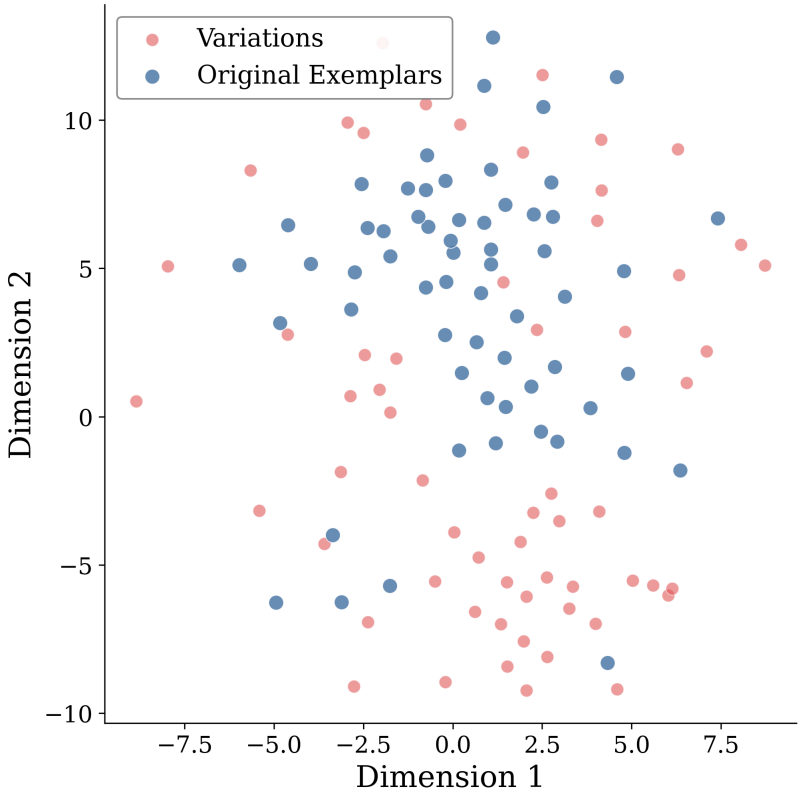
Notes: This figure shows the selection rates in resemblance to the source logo of AI variation when presented against the most similar logo from the same contest.

Study 2. We investigated whether AI-variations are less visually homogenized than the original exemplars in the Armor Health experiment. We use CLIP embeddings to represent

the logo's visual form. As an initial visualization, we project the embeddings into two dimensions using t-SNE (Van der Maaten and Hinton, 2008) in Figure WA-B.4. The variations cover a similar range as the exemplars, but are more dispersed, which is consistent with lower visual homogeneity.

To formalize this, we calculate each logo exemplar's embedding distance to its group centroid, where larger values indicate greater spread and thus higher visual heterogeneity. A paired t-test shows that variations exhibit substantially greater dispersion than their corresponding original exemplars ($Mean_{\text{Variation(1)}} = 0.068; Mean_{\text{Original}} = 0.054; t = 2.683; p = 0.008$).

Figure WA-B.4: 2-D Visualization of Forms of Exemplars and Variations



Notes: This figure shows the 2-D dimension reduced visualization of clip embeddings of exemplars and variations used in the Variation(1) condition. We first compute the principal components of embeddings, then conduct t-SNE. From the figure, while exemplars (blue) approximately represent the same range of variations (red), there is greater dispersion in variations and thus variations are less homogenized.

Study 3. To assess the effectiveness of the Optimization LoRA fine-tuning stage in en-

hancing a specific dimension of logo quality, in this case, click attractiveness, we compared variations generated with and without this optimization. Using the same 60 original logos from Study 1, we generated two sets of variations: one set using the pipeline with only the Logo LoRA fine-tuning, and another set using the full pipeline including the Optimization LoRA. This resulted in 60 pairs of logos: one with and one without Optimization LoRA; both logos are derived from the same input description.

We first notice distributional shifts in model outputs. For the 60 pairs, logos generated with Optimization LoRA exhibit higher levels of brightness ($\Delta Mean$ = 5.61, SE = 2.99, p = 0.065) and symmetry ($\Delta Mean$ = 0.0051, SE = 0.0024, p = 0.04). One illustrative example of the outputs without (on the left) and with Optimization LoRA (on the right) is in Figure WA-B.5. We can see that Optimization LoRA does not significantly change the logo rendering, but makes minor perturbations on features positively related to higher click attractiveness.

Table WA-B.4: Differences in Visual Features between Well-Performing and Ill-Performing Logos

Feature	**Mean difference (SE)**
Chromatic contrast	5.22 (5.80)
Luminance contrast	0.35 (1.24)
Colourfulness	0.03 (0.03)
Brightness	6.67** (3.45)
Saturation	4.45 (3.93)
Visual complexity	0.0008 (0.0014)
Horizontal symmetry	0.012* (0.0045)
Hue diversity	−0.52 (1.11)
White–background share	0.021 (0.027)

Notes: This table shows the differences in logo dimensions across the well-performing and ill-performing groups. Standard errors are in parentheses. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

We then study whether visual changes due to fine-tuning indeed lead to higher click at-

Figure WA-B.5: Logos Generated Using the Same Prompt without (Left Logo) and with (Right Logo) Optimization LoRA

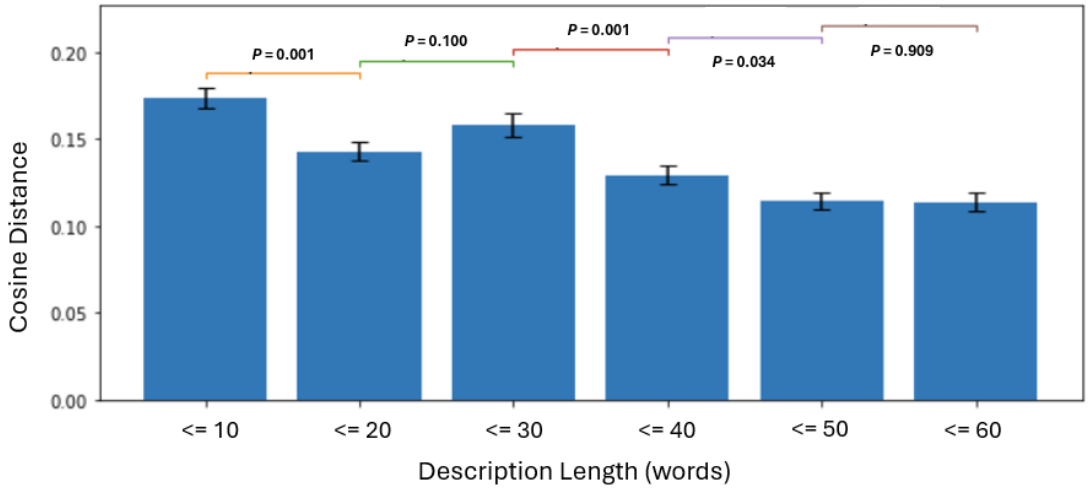
Notes: This figure shows variations generated by the model with and without the Optimization LoRA. The right logo exhibits higher level of symmetry and uses shades of green that are of higher brightness.

tractiveness. Similar to the data collection for training the Optimization LoRA, participants in an online survey were shown these pairs and asked to select the logo they were more likely to click on. Each pair was evaluated approximately 100 times. The mean selection rate for variations generated with the Optimization LoRA was 0.579 (t=5.075, significantly different from 0.5 with $p < 0.01$). This suggests that even with a relatively small labeled dataset (50 pairs for training), the contrastive fine-tuning process effectively guided the model towards producing outputs with enhanced performance on the targeted quality dimension.

B.4 Description Length to Control Variations

In this section, we investigate how the length of descriptions influences the perceptual similarity of variations to original exemplars. We inject system prompts to the image-to-text model to limit the length of the descriptions. We test descriptions of varying lengths, from 10 words to 50 words, generate variations based on these descriptions, calculate the cosine distances between variations and original exemplars, and present the results in Figure WA-B.6. As length increases, we observe an overall decreasing trend in cosine distances between variations and exemplars, meaning that as the descriptions become richer in information, the variations are more perceptually similar to the exemplars.

Figure WA-B.6: Perceptual Similarity between Exemplars and Variations from Descriptions of Varying Lengths



Notes: This figure shows the perceptual distance of AI-variations to the original logos. The variations are generated from descriptions of different lengths: (from left to right) 10, 15, 20, 30, 40, and 50. The y-axis represents the average cosine distance between the variations and the original logos. The heights of bars represent group means and the bounds represent one standard error. The p-values are from paired t-tests on variations generated by descriptions of consecutive lengths (e.g. 15 v.s 20; 20 v.s 30).

Description length provides only limited control over variation similarity. We observe no difference in exemplar distance between variations generated from 40-word and 50-word descriptions. This likely reflects two factors: logos are relatively low in visual complexity, so most of their information can already be captured in short descriptions; the image captioning model is not specifically trained to convey additional information as output length increases. We leave finer-grained control over variation similarity and its effects on designer outcomes for future research.

Web Appendix C. Robustness Checks on Experiment Results

In this section we provide additional robustness checks for our results in Section 4. We provide the full table of experimental results using the main measures discussed in the text: quality (Table WA-C.1) and originality (Table WA-C.2). For quality, we show in Table WA-C.1 that our results are robust for the mean quality and selected quantiles in Figure 8 (Top 5%, 10%, 25%, and 50%). For embedding-based originality, we show in Table WA-C.2 that our results are robust to the definition of 'high quality' by considering not only the top 50 in each condition, but also the top 100 in each condition as well as the top 200 or 400 in all submissions. Tables WA-C.3 and WA-C.4 present the contrasts of condition estimates of quality and embedding-based originality, respectively.

Table WA-C.1: Mean and Quantile Regression Results for Submission Quality

	Mean	$\tau = 0.95$	$\tau = 0.90$	$\tau = 0.75$	$\tau = 0.50$
Open	0.5735***	0.8193***	0.7727***	0.6989***	0.5908***
	(0.022)	(0.019)	(0.016)	(0.015)	(0.019)
Blind	0.5054***	0.7823***	0.7375***	0.6541***	0.5277***
	(0.024)	(0.019)	(0.016)	(0.016)	(0.019)
Variation(1)	0.5718***	0.8093***	0.7567***	0.6953***	0.5893***
	(0.024)	(0.022)	(0.019)	(0.017)	(0.020)
SubmissionTime	-0.0005	0.0010	0.0011	-0.0005	-0.0005
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
(Pseudo) R-squared	0.0306	0.0153	0.0169	0.0148	0.0152
Observations	2199	2199	2199	2199	2199

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Notes: Results are based on all submissions in the experiment. The regression includes designer-level controls: overall reputation, professionalism, hire again, quality, number of jobs, number of reviews, and hourly rate. The first column shows the mean click attractiveness, and the second to fifth columns show the click attractiveness of top 5%, 10%, 25%, and 50% submissions respectively. The rows show the estimates of three condition factors under specification (1) and designer-level demographics used as controls. Standard errors are clustered on the designer level.

Table WA-C.2: Mean Regression Results for Submission Embedding-Based Originality

	Top 50 Each Condition	Top 100 Each Condition	Top 200 in All Submissions	Top 400 in All Submissions
Open	0.0373***	0.0468***	0.0390***	0.0445***
	(0.0064)	(0.0048)	(0.0055)	(0.0047)
Blind	0.0547***	0.0646***	0.0509***	0.0626***
	(0.0068)	(0.0052)	(0.0064)	(0.0050)
Variation(1)	0.0512***	0.0598***	0.0499***	0.0577***
	(0.0078)	(0.0052)	(0.0070)	(0.0050)
SubmissionTime	0.0002	-0.0001	-0.0001	-0.0000
	(0.0004)	(0.0003)	(0.0003)	(0.0003)
R-squared	0.225	0.180	0.248	0.174
Observations	200	400	200	400

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Notes: Results are based on high-quality submissions in the experiment. The regression includes designer-level controls: overall reputation, professionalism, hire again, quality, number of jobs, number of reviews, and hourly rate. The columns show the embedding-based originality of high-quality submissions under different definitions. The first column uses the Top 50 submissions of each condition ranked by their click attractiveness. The second column uses the Top 100 submissions of each condition. The third column uses the Top 200 submissions in all conditions. The fourth column uses the Top 400 submissions in all conditions. The rows show the estimates of three condition factors under specification (1) and designer-level demographics used as controls. Standard errors are clustered on the designer level.

Table WA-C.3: Contrasts of Condition Estimates on Quality

Model	$\beta_{\text{Open}} - \beta_{\text{Var}}$	$\beta_{\text{Open}} - \beta_{\text{Blind}}$	$\beta_{\text{Var}} - \beta_{\text{Blind}}$
Mean	0.0017 (0.0188)	0.0681** (0.0213)	0.0665*** (0.0197)
$\tau = 0.95$	0.0100 (0.0156)	0.0370* (0.0154)	0.0270* (0.0138)
$\tau = 0.90$	0.0160 (0.0135)	0.0352** (0.0129)	0.0192 (0.0120)
$\tau = 0.75$	0.0036 (0.0125)	0.0448*** (0.0122)	0.0412*** (0.0115)
$\tau = 0.50$	0.0015 (0.0146)	0.0631*** (0.0149)	0.0616*** (0.0139)

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Notes: Results are based on estimates in Table WA-C.1. From left to right, the columns show the difference between the Open and the Variation conditions; the Open and the Blind conditions; the Variation and the Blind conditions. From top to bottom, the rows show the estimated effects on click attractiveness for Top 5%, 10%, 25%, and 50% submissions, respectively.

Table WA-C.4: Contrasts of Condition Estimates on Embedding-Based Originality

	$\beta_{\text{Open}} - \beta_{\text{Var}}$	$\beta_{\text{Open}} - \beta_{\text{Blind}}$	$\beta_{\text{Var}} - \beta_{\text{Blind}}$
Top 50 Each Condition	−0.0139** (0.0053)	−0.0174*** (0.0052)	−0.0035 (0.0054)
Top 100 Each Condition	−0.0130** (0.0040)	−0.0178*** (0.0046)	−0.0049 (0.0045)
Top 200 in All Submissions	−0.0110* (0.0051)	−0.0119* (0.0056)	−0.0009 (0.0054)
Top 400 in All Submissions	−0.0133*** (0.0038)	−0.0181*** (0.0047)	−0.0048 (0.0042)

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Notes: Results are based on estimates in Table WA-C.2. From left to right, the columns show the difference between the Open and the Variation conditions; the Open and the Blind conditions; the Variation and the Blind conditions. From top to bottom, the rows show the estimated effects on embedding-based originality for the four definitions of high-quality submissions, respectively.

Web Appendix D. Results of Additional Application

This section presents the analysis of the additional application described in Section 5. In Figure WA-D.1, we show the brief that designers observed in the additional application. The brief is similar to the one in the main study, except for the focal firm and requested logo.

Figure WA-D.1: Brief for Study 2

Brand name: Juneberry

Juneberry is an organic, fresh, and healthy vegan café. We have high-quality products and want to convey this quality in our logo.

Keywords: fresh, quality, vibrant, light, nutritious, nourishing, real food, premium, sustainable, fun

About the logo:
We want the logo to be modern and simple. We are open to ideas and any color scheme but envision more muted/lighter colors. Please include the text 'Juneberry' in the logo.
When submitting your designs, please use a white background and do not apply any special rendering effects.

IMPORTANT: We want to have a logo that can make our Facebook ads **more engaging and attract more clicks**.

We look forward to reviewing your creative submissions!

D.1 Participation

Table WA-D.1 shows the summary statistics of the designer-level variables of participating designers across the three conditions. We conduct a randomization check by doing t-tests for each variable separately and confirm that there are no statistically significant differences in any of these variables across conditions.

D.2 Results

Table WA-D.3 shows the average treatment effect estimates on submission quality. As in the main study, we use click attractiveness as the quality measure and collect it using the

Table WA-D.1: Designer-Level Summary Statistics

	Open	Variation	Blind
Reputation	4.12 (1.77)	4.09 (1.86)	4.05 (1.82)
Quality	4.13 (1.77)	4.10 (1.87)	4.04 (1.83)
Professionalism	4.13 (1.77)	4.09 (1.86)	4.07 (1.83)
HireAgain	4.11 (1.76)	4.10 (1.87)	4.03 (1.83)
NumJobs	22.05 (34.92)	20.16 (32.06)	18.58 (25.12)
Reviews	21.45 (34.16)	19.53 (31.23)	17.86 (24.07)
HourlyRate	21.02 (22.10)	24.31 (33.00)	22.43 (19.92)

Notes: This table shows the designer-level demographics across the three conditions. All differences are not significant. Standard deviations are in parentheses.

same method as in the main study. Table WA-D.5 shows the contrasts between condition estimates for selected quantiles and Figure WA-D.2 shows the differences on the full support of quantiles. All results are perfectly aligned with the main study.

Table WA-D.2: Mean Regression Results for Submission Embedding-Based Originality

	Top 50 Each Condition	**Top 100 Each Condition**	**Top 200 All Conditions**	**Top 400 All Conditions**
Open	0.1429***	0.0997***	0.1232***	0.0942***
	(0.0286)	(0.0229)	(0.0229)	(0.0185)
Variation(4)	0.1803***	0.1393***	0.1587***	0.1346***
	(0.0236)	(0.0194)	(0.0193)	(0.0169)
Blind	0.1887***	0.1361***	0.1733***	0.1343***
	(0.0280)	(0.0224)	(0.0246)	(0.0195)
SubmissionTime	-0.0036	-0.0000	-0.0018	0.0004
	(0.0024)	(0.0019)	(0.0019)	(0.0017)
$\mathbf{R}^2$	0.2942	0.1637	0.2888	0.1654
Observations	150	300	200	400

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Notes: Results are based on high-quality submissions in the experiment. The regression includes designer-level controls: overall reputation, professionalism, hire again, quality, number of jobs, number of reviews, and hourly rate. The columns show the embedding-based originality of high-quality submissions under different definitions. The first column uses the Top 50 submissions of each condition ranked by their click attractiveness. The second column uses the Top 100 submissions of each condition. The third column uses the Top 200 submissions in all conditions. The fourth column uses the Top 400 submissions in all conditions. The rows show the estimates of three condition factors under specification (1) and designer-level demographics used as controls. Standard errors are clustered on the designer level.

Table WA-D.3: Mean and Quantile Regression Results for Submission Quality

	Mean	$\tau = 0.95$	$\tau = 0.90$	$\tau = 0.75$	$\tau = 0.50$
Open	0.4965***	0.7371***	0.7011***	0.6140***	0.5203***
	(0.0415)	(0.0330)	(0.0303)	(0.0286)	(0.0289)
Variation(4)	0.4774***	0.7443***	0.7016***	0.5923***	0.4876***
	(0.0414)	(0.0349)	(0.0324)	(0.0303)	(0.0292)
Blind	0.4251***	0.6910***	0.6500***	0.5568***	0.4519***
	(0.0384)	(0.0348)	(0.0315)	(0.0292)	(0.0290)
SubmissionTime	0.0062	0.0088**	0.0080*	0.0078*	0.0058
	(0.0035)	(0.0034)	(0.0034)	(0.0033)	(0.0032)
(Pseudo) R^2	0.0516	0.0477	0.0420	0.0333	0.0268
Observations	1027	1027	1027	1027	1027

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Notes: Results are based on all submissions in the experiment. The regression includes designer-level controls: overall reputation, professionalism, hire again, quality, number of jobs, number of reviews, and hourly rate. The first column shows the mean click attractiveness, and the second to fifth columns show the click attractiveness of top 5%, 10%, 25%, and 50% submissions respectively. The rows show the estimates of three condition factors under specification (1) and designer-level demographics used as controls. Standard errors are clustered on the designer level.

Table WA-D.4: Contrasts of Condition Estimates on Embedding-based Originality

	$\beta_{\text{Open}} - \beta_{\text{Variation}}$	$\beta_{\text{Open}} - \beta_{\text{Blind}}$	$\beta_{\text{Variation}} - \beta_{\text{Blind}}$
Top 50 Each Condition	−0.0374*	−0.0457**	−0.0084
	(0.0181)	(0.0165)	(0.0176)
Top 100 Each Condition	−0.0396**	−0.0364**	0.0032
	(0.0142)	(0.0129)	(0.0136)
Top 200 in All Conditions	−0.0355*	−0.0501***	−0.0145
	(0.0148)	(0.0139)	(0.0153)
Top 400 in All Conditions	−0.0404**	−0.0401***	0.0003
	(0.0123)	(0.0120)	(0.0125)

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Notes: Results are based on estimates in Table WA-D.2. From left to right, the columns show the difference between the Open and the Variation conditions; the Open and the Blind conditions; and the Variation and the Blind conditions. From top to bottom, the rows show the estimated effects on embedding-based originality for the four definitions of high-quality submissions, respectively.

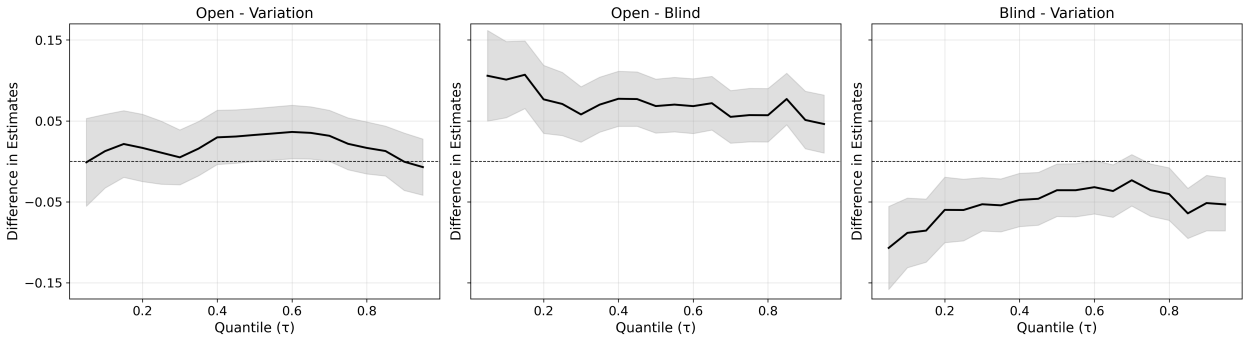
Table WA-D.5: Contrasts of Condition Estimates on Quality

Model	$\beta_{\text{Open}} - \beta_{\text{Variation}}$	$\beta_{\text{Open}} - \beta_{\text{Blind}}$	$\beta_{\text{Variation}} - \beta_{\text{Blind}}$
Mean	0.0191	0.0714**	0.0523*
	(0.0246)	(0.0259)	(0.0265)
$\tau = 0.95$	−0.0071	0.0461*	0.0532**
	(0.0177)	(0.0183)	(0.0167)
$\tau = 0.90$	−0.0005	0.0511**	0.0516**
	(0.0181)	(0.0181)	(0.0175)
$\tau = 0.75$	0.0217	0.0571***	0.0355*
	(0.0163)	(0.0168)	(0.0165)
$\tau = 0.50$	0.0326	0.0683***	0.0357*
	(0.0167)	(0.0170)	(0.0165)

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Notes: Results are based on estimates in Table WA-D.3. From left to right, the columns show the difference between the Open and the Variation conditions; the Open and the Blind conditions; and the Variation and the Blind conditions. From top to bottom, the rows show the estimated effects on mean click attractiveness, followed by the top 5%, 10%, 25%, and 50% submissions respectively.

Figure WA-D.2: Contrasts of Estimates on Submission Quality between Conditions



Notes: This figure shows contrasts of condition estimates with 95% CI. The panels show pairwise differences in click attractiveness estimates across quantiles: Open vs Variation (left panel); Open vs Blind (middle panel); and Blind vs Variation (right panel). The estimates are from quantile regression under specification (1), with click attractiveness being the dependent variable. On each panel, the y-axis shows the difference in estimates, and the x-axis shows the quantile level(τ).