Leveraging Generative AI to Create Visual Content in Digital Advertising

Remi Daviet[1] and Yohei Nishimura[1]

[1]Department of Marketing, Wisconsin School of Business, University of Wisconsin–Madison

April 2, 2026

Abstract

Generative AI for image synthesis has the potential to transform the digital advertising industry. However, a wide range of uncertainties persists regarding its integration into traditional advertising processes, including finding effective implementations, training methodologies, and achievable performance gains. Specifically, two core challenges limit its practical adoption: a search problem of finding high-performing visuals in a vast creative space, and an alignment problem of ensuring brand and campaign compatibility. This paper proposes a novel end-to-end framework that combines a generative AI with two predictive Bayesian neural networks to identify high-performance and brand-acceptable visuals. We develop a cost-effective Bayesian active learning approach solving simultaneously the dual objectives of performance and alignment. We test the framework in a live advertising campaign for an outdoor activities company. Our system generated a portfolio of visuals achieving higher mean click-through rate and more consistency (lower variance) than creatives from both a professional human designer and a competing AI model optimizing purely for aesthetics. This research provides a validated methodology that bridges the gap between the theoretical potential of generative AI and its practical application, offering a cost-effective solution to the critical search and alignment problems in creative design.

1 Introduction

In the competitive landscape of digital advertising, the success of a creative hinges on the quality of its components, with visual assets (e.g., images, illustrations, or backgrounds) often serving as a cornerstone for capturing consumer attention and conveying a message. The traditional process of producing these visual assets is a resource-intensive endeavor. Diverse specialist teams need to produce imagery in a process that can range from editing stock photos to organizing custom photoshoots. As a result, this process is often characterized by high production costs while facing considerable uncertainty regarding the eventual performance of the final ad. Consequently, unless a firm possesses substantial marketing resources, only a handful of visual concepts can be produced and tested. This inherent limitation in asset production constrains the ability to explore the vast space of creative possibilities, leaving significant potential for performance improvement untapped.

The advent of Generative Artificial Intelligence (AI) presents a paradigm shift, offering a compelling solution to these long-standing constraints by generating a multitude of design variations at a fraction of the time and cost of manual methods. However, this transition from technological promise to a practical advertising tool introduces a new set of managerial challenges. While some commercial "black-box" platforms are being developed, their proprietary methodologies offer little transparency or academic insight. Moreover, most of these generative models do not optimize for a given campaign or product. For practitioners and scholars, a critical gap remains: there is no established, scientifically vetted framework for systematically navigating the complexities of AI-driven creative optimization. In essence, generative AI shifts the core managerial challenge from the costly *production* of a few creatives to the strategic *navigation* of a vast creative space. This new paradigm introduces two fundamental problems. The first is a search problem: the latent "embedding" space of a modern generative model is immensely high-dimensional, making it prohibitively expensive to find the rare visual configurations that will resonate with consumers. The second is an alignment problem: generated visuals must also be "brand-acceptable". Specifically, they need to adhere to brand guidelines, avoid unrealistic artifacts, and align with campaign objectives.

This paper proposes and validates a novel framework that directly addresses these challenges by integrating a generative AI pipeline with a dual-objective Bayesian active learning system to efficiently discover high-performing and brand-acceptable visual content. We develop an end-to-end system that automates the creative process from visual generation to in-field performance measurement, providing a transparent and replicable methodology for AI-powered advertising.

The core of our solution is the development of a novel Bayesian active learning approach to simultaneously solve the coupled search and alignment problems. We design two predictive Bayesian neural networks: one to predict ad performance ($PerfAI$) and another to predict brand acceptability ($AcceptAI$). These networks intelligently guide the exploration of the generative model's high-dimensional embedding space. This active search methodology drastically reduces the amount of costly field data and manual vetting required for training. By focusing learning on the most informative creative variations, our approach makes systematic creative optimization feasible even for firms with limited resources, a significant departure from resource-intensive A/B testing, especially when the testing space is large.

We demonstrate the real-world viability of our framework through a field experiment with an outdoor activities company. In a live Instagram campaign, we tasked our system with generating background images for ads promoting nature exploration activities. The results show that our system can reliably generate a diverse set of visuals that achieve a mean click-through rate comparable to, and with greater consistency than, (1) creatives produced by a professional human designer and (2) a separate set of AI-generated visuals optimized purely for aesthetic appeal.

This paper makes several contributions to the marketing literature and practice. We provide one of the first end-to-end frameworks for leveraging generative AI in a live digital advertising setting. Methodologically, we introduce a cost-effective Bayesian active learning approach to solve the critical dual challenge of optimizing for both performance and brand alignment in a high-dimensional creative space. Empirically, we offer a proof-of-concept that synthesized content can compete with human-designed content, establishing a lower bound on the performance achievable with such systems. The managerial impact is significant: our framework provides a blueprint for a new class of creative optimization systems that can augment human designers, accelerate testing, and unlock performance gains, ultimately bridging the gap between the theoretical promise of generative AI and its effective, practical implementation.

2 Literature Review and Problem Definition

Our research is at the intersection of three rapidly evolving domains: generative AI for visual content creation, adaptive experimentation for creative optimization, and Bayesian optimization in high-dimensional spaces. This section reviews the key literature in each area to identify critical gaps and formally define the methodological problem that our framework is designed to solve.

2.1 Generative AI for Visual Content Creation in Marketing

The application of artificial intelligence to visual analytics has gained significant traction in marketing, enabling firms to extract insights from visual data for a range of predictive tasks (Dzyabura et al., 2023; Liu et al., 2020). More recently, the focus has shifted from analysis to synthesis, with generative models offering the potential to create novel visual content. Foundational work in this area has demonstrated how generative models can be used to understand consumer perceptions and preferences for product aesthetics (Burnap et al., 2023; Sisodia et al., 2024). For example, Burnap et al. (2023) use generative models to augment the aesthetic design process in the automotive industry, showcasing its potential for cost savings and creative exploration.

While these studies establish the viability of using AI to generate managerially relevant visual content, their application has been primarily in contexts of *post-hoc* analysis or controlled experiments. A common thread is the reliance on large, pre-existing labeled datasets—such as thousands of aesthetic ratings or product choices—to train the underlying predictive models. This raises a critical question for real-world advertising: how can such models be trained effectively when each new data point corresponds to an expensive and

time-consuming field test of an ad creative? The "cold-start" problem of learning what works from a near-zero base of field data for a specific campaign remains a significant, unsolved issue.

Our work addresses this critical gap by introducing a cost-effective Bayesian active learning approach. Instead of requiring a large, pre-labeled dataset, our system is designed to learn efficiently from small, sequentially tested batches of creatives.

2.2 Active Learning and Adaptive Experimentation for Creative Optimization

Active learning methods focus on iteratively improving model performance by adaptively querying the most informative data points, thereby reducing experimentation or data acquisition costs. In marketing, these techniques have been successfully applied to efficiently elicit consumer preferences in challenging applications (Dzyabura and Hauser, 2011; Huang and Luo, 2016), including measuring preferences over unstructured data, such as in cold-start recommendation problems (Dew, 2024).

In the specific context of digital advertising, the challenge of optimizing creatives has been extensively studied within the adaptive experimentation literature. The dominant methodologies, such as A/B testing and more sophisticated multi-armed bandit (MAB) algorithms like Thompson Sampling, have proven highly effective for allocating advertising budget among a *pre-defined, finite set* of creatives to maximize performance (Schwartz et al., 2017; Feit and Berman, 2019). These methods excel at balancing the exploration-exploitation trade-off to efficiently identify the best-performing creative from a given set.

However, a core limitation of these approaches is that they are fundamentally designed for *selection*, not *generation*. They provide a powerful answer to the question, "Which of these five options is the best?" but are unequipped to answer the more strategic question, "What option should I create next?". With the advent of generative AI, the space of potential creatives is no longer a small fixed set but a high-dimensional continuum. This creates a disconnect: the methods for testing creatives assume a small discrete set of options, while the technology for generating them offers infinite possibilities.

Our framework bridges this divide by integrating the creative generation process directly into the active learning loop. We leverage the principles of active learning not just to select from a fixed set, but to guide the generative AI toward more promising regions of the vast creative space. By unifying generation and testing, our approach systematically explores this space, revealing consumer preferences through field experiments and connecting the two previously disconnected stages of the creative process.

2.3 Bayesian Optimization in High-Dimensional Spaces

The methodological foundation of our approach is Bayesian optimization, a powerful sample-efficient technique for optimizing black-box functions where evaluations are expensive (Brochu et al., 2010; Shahriari et al., 2015). This framework is well-suited to our problem, where each function evaluation corresponds to a costly field experiment. However, a primary technical challenge is the "curse of dimensionality." Standard Bayesian optimization, particularly when

using Gaussian Process surrogates, struggles to scale to high-dimensional problems like navigating the latent space of a generative model. Most advanced Bayesian optimization libraries, such as BoTorch, are effective in spaces of a few dozens of dimensions at most (Balandat et al., 2020) .

One approach to address the high-dimensionality challenge is through methods like unsupervised dimensionality reduction (e.g., Principal Component analysis). These techniques might however struggle to effectively reduce an embedding space from a generative AI as most of these models are trained to have uncorrelated and incompressible embeddings. Moreover, these techniques do not guarantee that the most important features for predicting performance are preserved as dimensions are reduced. Finally, a critical and largely unaddressed gap in the marketing literature is how to perform this optimization under multiple, often conflicting, objectives. In advertising, it is not sufficient to find a high-performing creative: that creative must also align with brand standards and campaign goals.

To address this, we employ Bayesian neural networks (BNNs) as flexible surrogate models for both performance and brand acceptability. The use of a surrogate model is necessary to create a computationally tractable map of the relationship between creative inputs and performance outcomes. BNNs are particularly advantageous for this task. First, their neural network architecture is highly flexible, making them well-suited to capture the complex nonlinear relationships between visual embeddings and consumer responses that simpler models would miss. Second, the Bayesian framework provides effective uncertainty quantification (Izmailov et al., 2021). Indeed, BNNs do not just yield point predictions: they produce a full posterior distribution over outcomes, which naturally quantifies the model's confidence. This is crucial for active learning, as the model's uncertainty is the primary signal used to guide an efficient, information-driven search of the vast creative space.

Our key methodological contribution, however, goes beyond the choice of a Bayesian surrogate model and lies in the formalization of the creative optimization task as a *dual-objective search problem*. By explicitly modeling and optimizing for both performance and brand acceptability, our system ensures that the exploration remains constrained within a managerially relevant and brand-safe subspace. This approach is a novel application within the context of generative advertising and directly solves a critical practical challenge for managers seeking to adopt these new technologies responsibly.

3 Method

3.1 Application Context and Framework Overview

The effectiveness of our framework is demonstrated through an application conducted in partnership with an outdoor activities company. The firm's central challenge is to generate high-performing background images for its digital advertising campaigns. The primary objectives are twofold: (1) to maximize the advertising effectiveness, measured by the click-through rate (CTR), and (2) to ensure that all generated visuals are brand-acceptable. In this context, acceptability required that the images be realistic and representative of the natural landscapes relevant to the company's service areas. This dual-objective problem is common in marketing but is amplified by the capabilities of generative AI, which can produce

a near-infinite variety of visual content, making a systematic search for optimal creatives a significant challenge.

To address this, we formally define three core concepts that guide our methodology. *Performance* refers to the predicted in-field effectiveness of a visual, quantified in our study by its CTR. *Acceptability* is a measure indicating whether a visual adheres to the partner firm's brand standards and campaign requirements. Finally, the *Creative Space* is the high-dimensional embedding space defined by a generative model, representing the universe of all possible visual outputs that can be created.

The system we develop consists of three primary components designed to navigate this creative space efficiently. The first is an application-specific Generative AI (GenAI), which is custom-built to produce high-quality landscape images. The second component comprises two predictive models, AcceptAI and PerfAI, which are trained to evaluate generated visuals on the dimensions of acceptability and performance, respectively. The third and final component is a Bayesian active learning engine, which serves as the core of our system by intelligently selecting small informative batches of creatives for field testing to train the predictive models with maximum data efficiency.

While the generative and predictive architectures are necessarily tailored to the specific task at hand, they are modular by design and can be substituted with alternative models as technology evolves. Consequently, our primary methodological contribution lies in the integrated end-to-end framework and the Bayesian active learning procedure that governs the interaction between these parts to solve a dual-objective search problem in high-dimensional spaces. This system is adaptable to a wide range of creative domains beyond visual advertising, providing a robust and cost-effective solution for firms seeking to leverage generative AI for creative optimization, even with limited resources for experimentation.

3.2 Application-Specific Component: Generative Model (GenAI)

The generative model is a modular component of our framework that defines the navigable creative space to be explored. The choice of a specific GenAI architecture is application-dependent. Generative models that are highly specialized in a single domain can often compress visual information more efficiently, resulting in a creative space with a more manageable dimensionality, which is a highly desirable property for subsequent optimization tasks. There exist several architectures with pretrained models available off-the-shelf. For instance, the StyleGAN architecture (Karras et al., 2019) has readily available models capable of generating realistic faces, cars, rooms, or animals from a 512-dimensional creative space. In contrast, general-purpose models often require much higher-dimensional embeddings to capture a wider variety of concepts, posing a significant challenge for systematic search.

The key requirement for the generative component in our framework is that it provides a smooth latent embedding space of moderate dimensionality over which the Bayesian optimization procedure can operate effectively. The VAE-GAN architecture described below satisfies this by construction: its encoder maps images to a compact, continuous latent space (48 dimensions for layout, 256 for style) where nearby points are semantically similar and interpolation yields smooth transitions, an essential property for gradient-based active learning. By contrast, diffusion models and off-the-shelf pre-trained foundation models often

operate in substantially higher-dimensional spaces that are typically "uncorrelated and incompressible," posing a significant curse of dimensionality challenge for systematic Bayesian optimization.

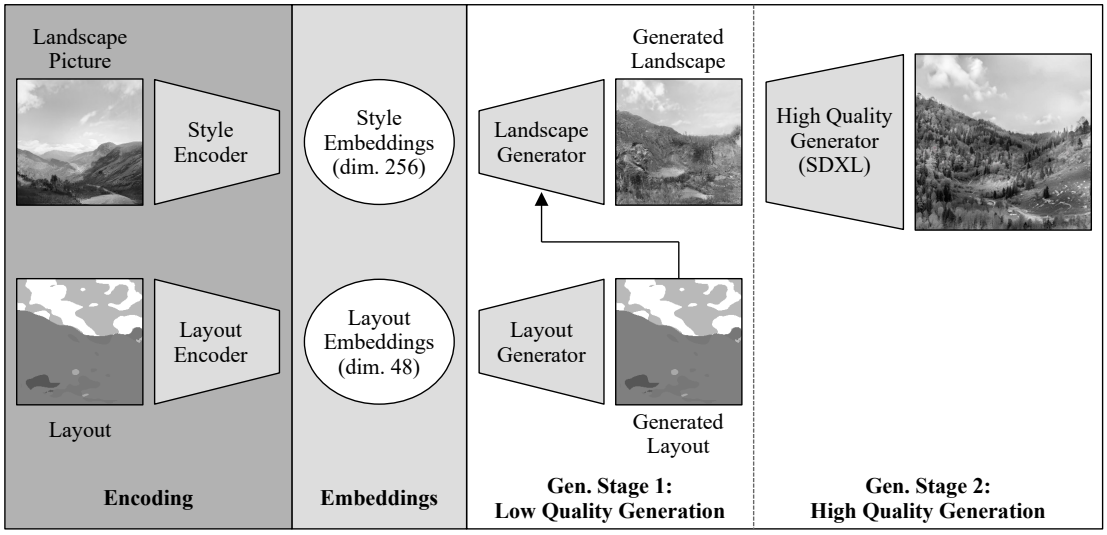


Figure 1: The two-stage architecture of the generative model (GenAI). Stage 1 generates a low-resolution image from separate layout and style embeddings. Stage 2 enhances this image to a high-quality final output.

We navigate this trade-off between searchability and visual quality by adopting a custom two-stage approach: a first stage generates a low-resolution image and is engineered to define a 304-dimensional embedding space that is sufficiently expressive yet manageable for our active learning engine, while a high-fidelity diffusion model handles the final aesthetic enhancement.

As represented in Figure 1, the first stage disentangles the visual representation into two separate components. For the layout, we develop and train an adversarial variational autoencoder (VAE-GAN) that generates and encodes semantic layouts from a 48-dimensional *layout embedding*. This embedding controls the high-level structure of the scene, such as the location of sky, trees, and water. Details for the architecture, training dataset, and training procedure for our VAE-GAN model are provided in Online Appendix A. For the low-resolution image generation, we use the same dataset to train a separate 256-dimensional *style embedding* model, based on the SPADE architecture Park et al. (2019), that apply textural and atmospheric qualities to the layout (e.g., weather, time of day, type of trees).

This disentangled layout-style architecture provides an important practical advantage: it allows for manual intervention to correct visual artifacts. As the active learning process guides the GenAI to explore novel regions of the creative space where training data may be sparse, it can sometimes produce unrealistic images, such as a landscape with a tree erroneously placed in the sky. Our model's separation of layout and style enables a human vetter to directly edit the semantic layout to correct such flaws before the image is finalized. An example of this correction process is shown in Table 1. This correction facilitates the training of the AcceptAI by providing examples that might be extremely similar, but that

differ in acceptability. As a result, the AcceptAI can more efficiently learn the "frontier of acceptability" in the embedding space.

Table 1: Example of Manual Layout Correction

Layout		**Picture**	
Original	Corrected	Original	Corrected
			

The second stage of our GenAI focuses on enhancing the generated low-resolution image to a high quality suitable for a digital advertising campaign. This component is designed to be modular. In our implementation, we use a publicly available SDXL model (Podell et al., 2023), but any off-the-shelf solution capable of improving image quality can be substituted. This includes open-source models like the Stable Diffusion family and FLUX, or commercial models such as Google Gemini or OpenAI's GPT image generator. This modularity ensures that the framework can readily incorporate advances in image generation technology to continually improve the quality of the final creative outputs.

3.3 Predictive Models: Mapping the Creative Space

To navigate the 304-dimensional creative space defined by the GenAI, we employ two Bayesian neural networks (BNNs) that act as surrogate models. These models map a location in the creative space to its predicted acceptability and performance, respectively. Their structure is detailed in Online Appendix B.

3.3.1 The Acceptability Model (AcceptAI)

The primary role of AcceptAI is to serve as an efficient, automated filter that predicts brand alignment. This is crucial for preventing the system from searching for informative batches or best performers in the space of unacceptable images. Without it, generated batches would continuously be rejected by the brand for containing unacceptable images and field testing could not be done. In our application, this means filtering out images with visual artifacts (e.g., unrealistic elements) and landscapes not representative of the company's serviced geographical locations. This concept is highly extensible and could be adapted to a wide range of brand-specific criteria, such as emotional tone, camera point of view, or the shape of an object.

Formally, given a 304-dimensional embedding vector x from the GenAI's creative space, AcceptAI outputs the predicted probability of the corresponding visual being brand-acceptable:

$$P(\text{Acceptable}|x, \theta_A) = f_{\text{AcceptAI}}(x; \theta_A), \tag{1}$$

where θ_A are the parameters of a fully connected BNN.

3.3.2 The Performance Model (PerfAI)

The PerfAI model is designed to predict the in-field advertising performance of a given visual by creating a response surface over the creative space. This model maps the visual characteristics encoded in the embedding vector w to an expected Click-Through Rate (CTR). Formally, $E[\text{CTR}|x, \theta_P] = f_{\text{PerfAI}}(x; \theta_P)$, where θ_P are the parameters of a BNN with an identical architecture to AcceptAI, but with an output layer scaled to the expected range of CTRs (0 to 10% in our case). To account for potential temporal shifts in CTR during sequential field tests, the model also includes a series of learnable seasonal parameters in the form of batch fixed effects, as detailed in Online Appendix B.

3.3.3 Bayesian Updating

The use of a Bayesian framework is central to our approach. Instead of seeking a single point estimate for the neural network parameters θ, we embrace uncertainty by treating the parameters themselves as random variables. Our goal is to infer the posterior probability distribution of the parameters, $p(\theta|\mathcal{D})$, which represents our updated beliefs about θ after observing data $\mathcal{D}$ (acceptability ratings or CTRs). This updating is governed by Bayes' theorem:

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta). \tag{2}$$

Here, $p(\theta)$ is the *prior distribution*, which encodes our beliefs about the parameters before seeing any data. The term $p(\mathcal{D}|\theta)$ is the likelihood of observing the data given a specific set of parameters. The result, $p(\theta|\mathcal{D})$, is the *posterior distribution*, which synthesizes our prior beliefs with the evidence from the data. This posterior distribution is critical as it captures our model's uncertainty.

For complex models like BNNs, the posterior distribution is analytically intractable. We therefore approximate it by drawing samples using a Hamiltonian Sequential Monte Carlo algorithm (Burda and Daviet, 2022), a method particularly well-suited for the high-dimensional parameter spaces of neural networks, and with good representation of uncertainty (Izmailov et al., 2021). This process generates a set of parameter samples $\{\theta^{(s)}\}_{s=1}^{S}$ from the posterior distribution, where each sample represents a plausible model. This ensemble of models is what allows us to quantify uncertainty and guides the active learning engine.

3.4 Acquiring Training Data through Bayesian Active Learning

Training deep learning models typically requires thousands of labeled examples, but each label for AcceptAI requires valuable time from a company professional, and each data point for PerfAI requires a costly live ad campaign. Our active learning engine solves two problems

simultaneously: (1) it trains the predictive models with improved data efficiency, and (2) it actively searches for the regions of the creative space that contain top-performing, brand-acceptable visuals. The novelty lies in its objective function, which unifies these goals. We posit that the models do not need to be accurate across the entire creative space, only in the regions where acceptable top performers are likely to be found. Therefore, our objective is to select batches of creatives that are maximally informative about the location of these high-value regions.

Let $\mathcal{X}_A$ be the (unknown) set of all embedding vectors x that correspond to brand-acceptable visuals, and $\mathcal{X}_P^*$ be the set of vectors corresponding to top-performing visuals. Our BNNs represent our beliefs (or uncertainty) about these sets. For instance, our current belief that a visual x is acceptable after observing data $\mathcal{D}$ is:

$$P(x \in \mathcal{X}_A|\mathcal{D}) = \int f_{\text{AcceptAI}}(x;\theta_A)p(\theta_A|\mathcal{D})d\theta_A, \tag{3}$$

and similarly, our current belief that a visual w could be a CTR maximizer is:

$$P(x \in \mathcal{X}_P^*|\mathcal{D}) = \int \mathbb{I}\left\{x \in \arg\max_{x'} f_{\text{PerfAI}}(x';\theta_P)\right\} p(\theta_P|\mathcal{D})d\theta_P, \tag{4}$$

where $\mathbb{I}$ is an indicator function equal to 1 if its content is true and 0 otherwise. In the discussion below, we will use the general notation $\mathcal{X}$ to indicate the optimal set when the approach is the same for the Acceptability and Performance prediction problems.

To learn about these sets efficiently, we use an approach based on maximizing the *Shannon information gain* (Chaloner and Verdinelli, 1995). As our beliefs about the BNN parameters θ are updated, so are our beliefs about where the best images are. After testing a new batch, our training dataset becomes $\mathcal{D}_{new}$ and we update our models, forming a more informed set of beliefs represented by the new posterior distribution. The Shannon information gain, measured by the Kullback-Leibler (KL) divergence, quantifies how much our beliefs changed:

$$D_{KL}\left(p(\mathcal{X}|\mathcal{D}_{new})\|p(\mathcal{X}|\mathcal{D}_{old})\right) = \int p(\mathcal{X}|\mathcal{D}_{new}) \log \frac{p(\mathcal{X}|\mathcal{D}_{new})}{p(\mathcal{X}|\mathcal{D}_{old})} d\mathcal{X}. \tag{5}$$

By selecting the batch that we expect will cause the biggest change, we ensure we learn as much as possible from each expensive experiment. This process represents a departure from the traditional Bayesian active learning approach, where the information gain is usually measured in terms of change in $p(\theta)$ instead of $p(\mathcal{X})$. This change is what allows us to focus the learning on the region of the creative space with the most potential, instead of learning to predict accurately over the whole creative space, representing a substantial efficiency gain. In what follows, we omit $\mathcal{D}_{old}$ from equations and only represent the change in dataset after an observation to simplify notation.

To train AcceptAI, we find a batch X^B with (unknown) acceptability ratings A^B that maximizes the expected information gain about the set of acceptable visuals $\mathcal{X}_A$:

$$\max_{X^B} \mathbb{E}_{A^B \sim p(A^B|X^B)} \left[D_{KL}\left(p(\mathcal{X}_A|A^B, X^B)\|p(\mathcal{X}_A)\right)\right] \tag{6}$$

$$= \max_{X^B} \underbrace{\int p(A^B|X^B)}_{\substack{\text{Average over all} \\ \text{probable outcomes}}} \underbrace{D_{KL}\left(p(\mathcal{X}_A|A^B, X^B)\|p(\mathcal{X}_A)\right)}_{\substack{\text{Information gained if a} \\ \text{specific outcome occurs}}} dA^B \tag{7}$$

Here, $p(\mathcal{X}_A)$ is our *prior* belief about which visuals are acceptable. The inner integral is the information gain we would achieve if we observed a specific set of ratings A^B for our test batch X^B. Since we do not know these ratings beforehand, we average this gain over all possible outcomes, weighted by the probability $p(A^B|X^B)$ that current model believes this outcome could occur.

To train PerfAI, we solve a dual-objective optimization problem. We search for a batch X^B that is maximally informative about the set of top-performing visuals, $\mathcal{X}_P^*$, conditioned on the probability that the batch will be deemed acceptable. This integrates brand alignment directly into the search process:

$$\max_{X^B} \mathbb{E}_{A^B,\mathrm{CTR}^B \sim p(A^B,\mathrm{CTR}^B|X^B)} \left[D_{KL} \left(p(\mathcal{X}_P^*|\mathrm{CTR}^B, X^B) \| p(\mathcal{X}_P^*) \right) \right] \tag{8}$$

$$= \max_{X^B} \underbrace{\iint p(A^B|X^B) p(\mathrm{CTR}^B|X^B)}_{\substack{\text{Average over all probable} \\ \text{outcomes (ratings \& CTRs)}}} \underbrace{D_{KL} \left(p(\mathcal{X}_P^*|\mathrm{CTR}^B, X^B) \| p(\mathcal{X}_P^*) \right)}_{\substack{\text{Information gained if a specific} \\ \text{acceptable outcome occurs}}} dA^B d\mathrm{CTR}^B \tag{9}$$

where the expectation is taken over the joint distribution of acceptability ratings A^B and CTRs for the batch. This ensures our search for high-performing visuals is concentrated within the parts of the creative space that are also brand-acceptable.

3.5 Computational Implementation and Process

The objective functions in Equations 7 and 9 are computationally intractable and are approximated using numerical integration methods. To computationally approximate the KL divergence, we use a technique based on importance sampling. Specifically, we identify the optimal creative (maximizer) for each plausible model in our Bayesian ensemble and use Kernel Density Estimation (KDE) to construct smooth density approximations of our prior and posterior beliefs. This allows us to estimate the expected information gain at representative sample points, providing a stable objective for the gradient-based batch selection. Further details on the computational implementation are provided in Online Appendix C.

The end-to-end workflow proceeds in two phases:

1. **Training AcceptAI.** The active learning engine repeatedly selects an informative batch of 12 visuals. A company representative labels images in the batch as unacceptable (0.1), barely acceptable (0.5), good (0.7), or great (0.9). We use a scale ranging from 0.1 to 0.9 instead of 0 to 1 for numerical stability, a technique known as label smoothing. These new labels are then used to update the AcceptAI model, which in turns informs the selection of the next batch. If some images are close to be acceptable but require minor corrections (e.g., removing a tree in the sky), this correction is done here following the process described in Section 3.2. As mentioned previously, this allows the model to learn the frontier of acceptability by seeing training examples that are similar but not equivalently acceptable. This cycle is repeated until accuracy is acceptable (60 batches of 12 visuals in our application).

2. **Training PerfAI.** Using the trained AcceptAI as a filter, the active learning engine selects an informative batch of 12 visuals (using Equation 9), to which is added the

visual predicted to perform best, as well as the image from the previous sample scoring the highest CTR (from trial 2 onward) to facilitate the training of the seasonality parameters (see Online Appendix B). A human manually vets this batch. In this phase, images with minor visual flaws are manually corrected (as shown in Table 1) and re-encoded to the creative space to obtain the updated x, without retraining the AcceptAI to save computational time. If some images are unacceptable and cannot be corrected, the acceptability labels from the batch are used to update AcceptAI and a new active learning batch is produced. The vetted batch is then run in a live ad campaign to collect CTR data, and the results are used to update the PerfAI model. In our application, this cycle is repeated until 9 fully acceptable batches are tested.

3.6 Framework Extensibility and Portability

Our framework is designed to be both generalizable to other marketing problems and adaptable to future technological advances. This extensibility stems from three key features of its design.

First, the framework is modular. Each component—the GenAI, the predictive models, and the active learning engine—can be independently updated or replaced. For instance, as more powerful, lower-dimensional generative models become available, they can be swapped in without altering the core logic of the active learning engine. This ensures the framework remains relevant and is not tied to a specific model like SDXL that may become obsolete. The modularity also allows for several AcceptAIs to be used to score for various factor. For instance, one AI could score Aesthetics and another AI could score emotional response, with training metrics collected by survey or experiments. There are also cases where a pre-existing, continuously differentiable scoring model might be available and could be used in place of our Bayesian AcceptAI model, allowing us to skip the acceptability training.

The optimizer currently leverages the differentiability of our Bayesian surrogate models and the continuous latent space of the GenAI to perform efficient gradient-based batch selection. If the image generator were replaced by a black-box API, the framework would require the upstream model to expose its latent embedding space to maintain this optimization logic. If such latent vectors were unavailable or not smooth (such as with quantized VAEs or architectures using discrete codebooks), the search would necessarily shift to gradient-free optimization methods (e.g., CMA-ES or evolutionary algorithms). While such methods can navigate high-dimensional spaces, they typically require significantly more iterations to converge. Consequently, the modularity of our framework is most effective when the substitute GenAI model provides a continuous and navigable latent manifold, which ensures that the search for high-performing creatives remains sample-efficient and avoids the issues inherent in purely discrete or non-differentiable design spaces.

Second, the brand acceptability filter training can potentially be enhanced and scaled using Vision-Language Models (VLMs). We tested this approach using a lightweight but powerful model, Qwen3-VL-8B-Instruct (Qwen Team, 2025). We first evaluated the model on an objective classification task: identifying whether an image contained no body of water (0), a river or stream (1), a pond or lake (2), or the sea (3). Despite some inherent ambiguous cases, such as distinguishing between river banks and shorelines, the model achieved a Quadratic Weighted Kappa (QWK) of 0.73, indicating good agreement with ground truth

(1.0 being perfect agreement). However, for the more subjective task of determining whether a visual was brand-acceptable according to the criteria in Section 5, the model failed to match the brand manager's judgment. Despite testing dozens of prompt variations, we were unable to achieve a QWK score above 0.05. Using a larger model, Gemini 3 Thinking, did not yield any improvement.

This difficulty suggests that current LLMs struggle to serve as a "digital twin" of a brand manager for nuanced content evaluation. This aligns with recent research on behavior prediction, which suggests that even when provided with sophisticated data about a target human, LLM predictive accuracy remains limited, often around 70% compared to 60% for chance (Toubia et al., 2025), and performs even more poorly on content evaluation tasks (Peng et al., 2026). The task is particularly difficult given that ratings are often subjective and that human labelers themselves often exhibit intra-rater variability. Our results suggest that while VLMs are effective for objective feature extraction, the manual training of an AcceptAI model via active learning remains the most reliable option for capturing potentially subjective idiosyncratic brand preferences.

Finally, the framework is adaptable to other creative tasks beyond visual advertising. The core idea of using an active learning engine to navigate a generative space guided by one or more predictive models is broadly applicable. The 'AcceptAI' and 'PerfAI' models can be replaced with models that score any desired attribute, such as predicted emotional response, perceived luxury, or alignment with specific customer preference vectors. This transforms the framework into a general-purpose tool for multi-objective creative optimization across a wide range of marketing problems, from product design to personalized content creation.

4 Simulation Study

4.1 Objectives of the Simulation Study

Before deploying our framework in the field, we conduct a simulation study to validate the efficacy of the proposed Bayesian active learning engine. The primary goal is to assess the methodology in a controlled environment where the ground-truth performance landscape is known, a condition impossible to satisfy in live experiments. This setting allows us to rigorously test the core feasibility of our approach, specifically its ability to navigate a high-dimensional creative space ($D = 304$) and identify regions of high performance with minimal data. To ensure the reliability of our findings and account for stochastic variations in the optimization process, all simulation results are evaluated and averaged across five independent simulation runs. A replication package is provided[1].

4.2 Simulation Design and Ground Truth

To mirror the challenging aspects of real-world optimization, we design a simulation featuring several brand acceptability regions and sparse high-performance outliers. A critical challenge in this context is the disconnect between the representation used by the generative model and the perceptual criteria used by consumers. While market responses are often driven by

[1] `https://osf.io/re8tm/overview?view_only=cd452084a57a4f98aeada34935176f14`

a lower-dimensional set of latent features, such as layout or style, the optimization engine must search within the full, entangled embedding space without access to these disentangled coordinates. To simulate this relationship, we construct a ground truth defined on a lower-dimensional projection unknown to the search algorithm. This design mimics the "black-box" nature of the link between real-world generative embeddings and consumer response.

We first use Parametric UMAP to learn a fixed non-linear projection Φ from the 304-dimensional creative space $\mathcal{X}$ to a 4-dimensional latent perceptual space $\mathcal{U}$. We project the embeddings corresponding to the layout and style separately, resulting in two interpretable 2-dimensional components: a layout subspace and a style subspace, such that for any embedding $x \in \mathcal{X}$, $\Phi(x) = [u_{\text{layout}}, u_{\text{style}}] \in \mathbb{R}^2 \times \mathbb{R}^2$. The visualization of this projected space is shown in Figure 2. It is important to note that the active learning optimization engine operates exclusively in the original 304-dimensional space and has no knowledge of this projection or the underlying 4-dimensional coordinates.

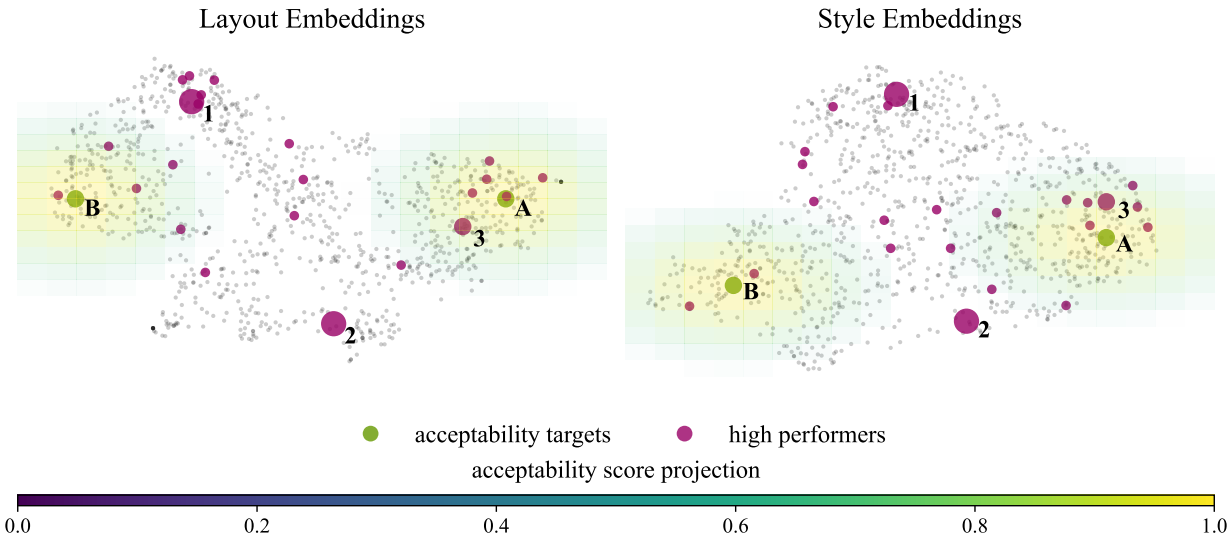


Figure 2: Projection of the 304-dimensional creative space into 2D Layout and 2D Style subspaces. The green regions indicate the brand-acceptable "islands," while the purple points represent the high-performance centroids. The optimizer searches in the 304D space, unaware of this underlying structure.

4.2.1 Brand Acceptability

Real-world brand guidelines often restrict acceptable creatives to specific, sometimes disconnected, regions of the design space (e.g., landscapes representing either mountains or seashore). To simulate this, we define two distinct "acceptability islands" in the projected space $\mathcal{U}$. The acceptability score, $S_A(x)$, is calculated using the Mean Squared Euclidean distance (MSE) between the projected point $\Phi(x)$ and two centroids, C_1^A and C_2^A. This metric, also called mean squared error, represents the average of the squared differences between the coordinates across the latent dimensions, providing a smooth measure of proximity to

the target brand concept. Formally:

$$S_A(x) = \max_{k \in \{1,2\}} \exp\left(-\frac{1}{\lambda}\text{MSE}(\Phi(x), C_k^A)\right) \tag{10}$$

where $\lambda = 10$ is the bandwidth parameter. The centroids are located at $C_1^A = [5.0, 1.5, 4.0, -1.0]$ and $C_2^A = [-5.0, 1.5, -4.0, -3.0]$ (green points in Figure 2). This multimodal definition challenges the active learning engine to explore diverse regions rather than converging on a single acceptability mode.

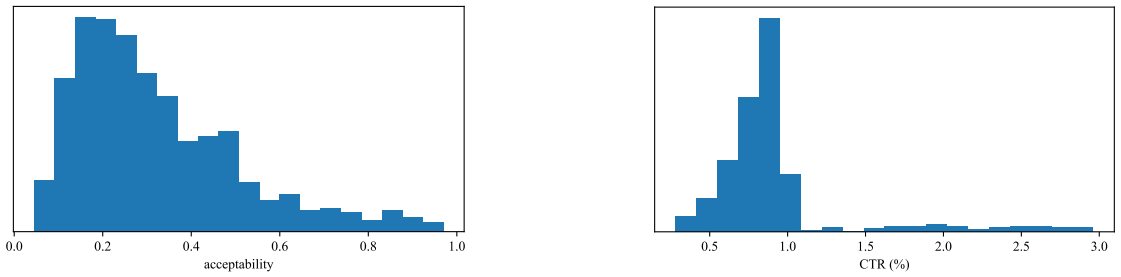
4.2.2 Performance Landscape

The performance landscape is modeled to be sparse and rugged, featuring isolated peaks of high performance. We define the expected Click-Through Rate (CTR), $\mu(x)$, based on the proximity of the projected point to a set of high-performance centroids in the latent space. Similar to the acceptability model, we utilize the Mean Squared Euclidean distance to map proximity to performance. These centroids include three "main" high-performance peaks with maximum achievable CTRs of 3%, 3%, and 2%, and ten randomly sampled "local optima" with a maximum CTR of 1%. The two peaks with 3% CTRs being outside of the acceptable area, the maximum achievable CTR within acceptable creatives is thus 2%. The expected CTR for any creative x is given by:

$$\mu(x) = \max_{j \in \{1 \ldots M\}} \beta_j \cdot \exp\left(-\frac{1}{\lambda}\text{MSE}(\Phi(x), C_j^P)\right) \tag{11}$$

where C_j^P is the location of the j-th performance peak, β_j is its maximum CTR, and $\lambda = 10$.

The resulting distributions of these ground truth scores across the creative space are illustrated in Figure 3. The long tails observed in both the acceptability and performance scores underscore the difficulty of the optimization task: high-performing and brand-acceptable creatives are rare. The search process must navigate a vast region of negligible value to discover the few high-value configurations.



(a) Distribution of Acceptability Scores

(b) Distribution of Expected CTRs

Figure 3: Distributions of ground truth scores across the creative space. High-performing and brand-acceptable creatives are rare, represented by the long tails of the distributions.

4.3 Training and Metrics

The simulation proceeds in two phases, (1) AcceptAI pre-training and (2) creative optimization including PerfAI training, mirroring the phased deployment strategy used in our field experiment. To ensure a robust evaluation across all phases of the simulation, we define three recurring metrics that apply to both the acceptability and performance dimensions:

1. **Peak Performance (x^*_{pred}):** The ground truth CTR of the single creative predicted to be the best by the model. This measures the system's ability to identify the global maximum of the performance landscape.

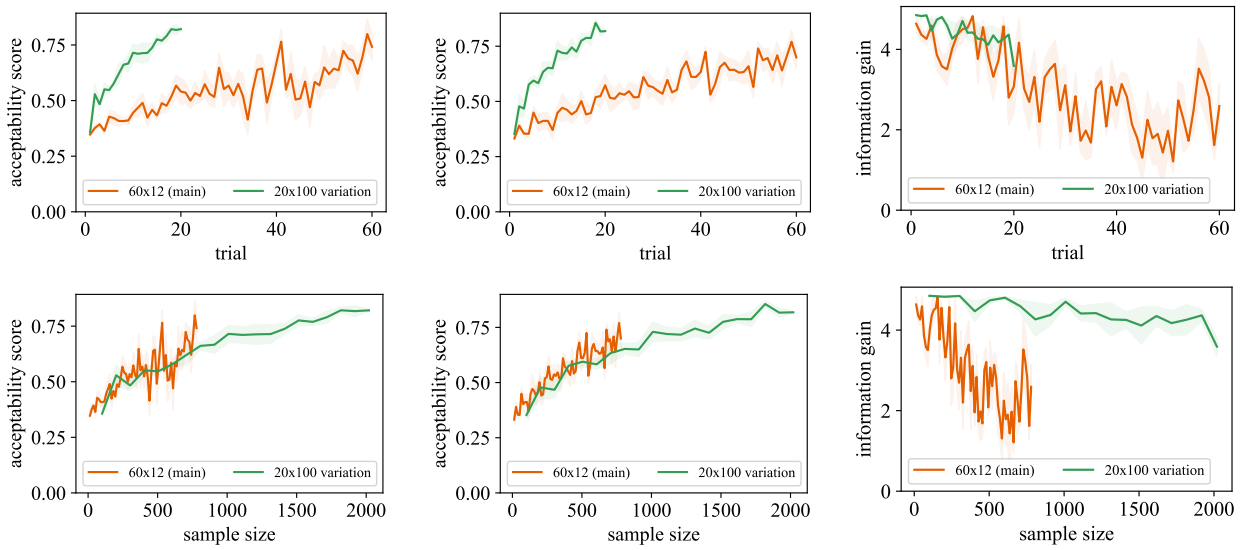
2. **Portfolio Quality (Probability Matching Average):** The average ground truth CTR of a batch of 10 creatives sampled via probability matching, where the sampling weight is proportional to the probability that the location is a maximizer (the mechanism used in Thompson Sampling). This metric ensures the model identifies a robust set of high-performing assets (a portfolio) rather than a single outlier, validating the calibration of the posterior distribution.

3. **Realized Information Gain (KL-Divergence):** The Kullback-Leibler divergence between the posterior distribution over the location of the maximizer before and after observing the batch results, as detailed in Online Appendix C.2. This metric directly quantifies the amount of knowledge gained about the optimal region from each experimental trial.

4.4 Phase 1: Brand Acceptability Pre-training

The first phase evaluates how efficiently the active learning engine can train the AcceptAI model to learn brand constraints. To evaluate the efficiency of the active learning engine, we compare two training configurations. In the first condition, we mirror the field experiment's design using 60 batches of 12 designs each (720 total training points). In the second condition, we use 20 batches of 100 designs (2,000 total training points). This comparison allows us to investigate whether more frequent model updates can compensate for a smaller total sample size. In both cases, the continuous scores from the ground truth acceptability function $S_A(x)$ are used as feedback to update the *AcceptAI* model sequentially.

The results from this phase (Figure 4) demonstrate that the active learning engine efficiently learns the brand acceptability frontier. We find that the higher update frequency in the 60-batch condition (720 points) allows the model to achieve a peak score and portfolio quality comparable to the 2,000-point condition, despite having only one-third of the total training data. The ability to re-orient the search after every 12 designs leads to a more targeted exploration of the acceptability islands, even if the information gained per batch is smaller. This suggests that frequent feedback cycles effectively compensate for a lower volume of manual labeling, making the iterative nature of active learning a powerful tool for firms with limited budgets.



(a) Predicted Best (x^*_{pred}) (b) Portfolio Quality (c) Realized Information Gain

Figure 4: Acceptability pre-training results per active learning trial (top) and same metrics per cumulative sample size (bottom). Frequent updates allow for high accuracy with significantly fewer manual labels. Shaded areas represent the standard error across 5 simulation runs.

4.5 Phase 2: Performance Optimization

In the second phase, we simulate the field testing process to train the PerfAI model. Each trial follows a specific protocol beginning with a simulation of human-in-the-loop refinement. For every generated batch, creatives with "near-miss" acceptability scores ($0.4 \leq S_A(x) < 0.5$) are flagged for correction. To mimic the manual adjustment of a design described in Section 3.2, we sample 1000 close neighbors around the original embedding x from a Gaussian distribution $\mathcal{N}(x, 0.1 \cdot I)$. The original creative is then replaced by the sampled neighbor x' that yields the highest acceptability score, provided it satisfies the acceptance condition $S_A(x') \geq 0.5$.

Following this correction step, the batch passes through an acceptability gate. If less than 75% of the designs in the corrected batch meet the acceptability threshold ($S_A(x) \geq 0.5$), the batch is rejected for field testing. In this scenario, the AcceptAI model is updated with the acceptability scores of the current batch (including corrected points) to better refine its boundary detection, and a new batch is subsequently requested. Conversely, if at least 75% of the designs are acceptable, the batch proceeds to testing, with any remaining unacceptable designs being removed from the set.

Finally, for the acceptable creatives that pass the gate, we generate synthetic field data. To realistically mimic the behavior of modern advertising platforms, we do not allocate impressions uniformly. Instead, the platform allocates more impressions to creatives with higher expected performance. We model this by distributing the total pool of impressions $N_{total} = 4,000 \times K$ among the K designs in the batch. The share of impressions w_i for a creative x_i is proportional to a power of its true expected CTR: $w_i \propto \mu(x_i)^\alpha$, with a

concentration parameter $\alpha = 0.02\sqrt{N_{total}}$. The specific number of impressions allocated to creative i is $n_i = N_{total} \cdot w_i$. These parameters have been selected to approximate impression numbers obtained in a small pilot experiment of random designs. As N_{total} becomes large, most of the impressions are allocated to top performers. The observed number of clicks is then drawn from a Binomial distribution, $Clicks_i \sim \text{Binomial}(n_i, \mu(x_i))$. This performance data is utilized to update the PerfAI model.

4.5.1 Benchmarks

We assess the performance of our framework by comparing it against two benchmarks and an ablation baseline, all evaluated across 5 independent simulation runs.

The first benchmark is a Random Policy, which serves as a baseline for the difficulty of the search problem. In each trial, a batch of 10 randomly selected acceptable creatives is tested. Additionally, the system also tests the single creative predicted to be the best by the current model, a strategy known as greedy optimization. This benchmark effectively measures whether intelligent search offers any advantage over random guessing within the acceptable regions.

The second benchmark is Thompson Sampling (TS), which represents a standard multi-armed bandit approach widely used in industry (Schwartz et al., 2017; Thompson, 1933). Instead of optimizing for information gain, this method samples a batch of 10 acceptable creatives from the posterior distribution given by the PerfAI model, with sampling weight proportional to the probability that the location is a maximizer. As with the random policy, the greedy strategy is also employed here by including the predicted best creative in the test set. This comparison isolates the value of the active learning component, specifically whether maximizing information gain leads to faster convergence than Thompson Sampling's probability matching approach.

Finally, we include an Ablation (No AcceptAI) condition to isolate the contribution of the brand-alignment component. In this scenario, the active learning engine optimizes for performance alone, without the guidance of the AcceptAI pre-training or filter. The greedy strategy is maintained, testing the predicted top performer alongside the optimized batch. This condition tests the hypothesis that pre-filtering for acceptability is more efficient than searching the entire space and discarding failures post-hoc.

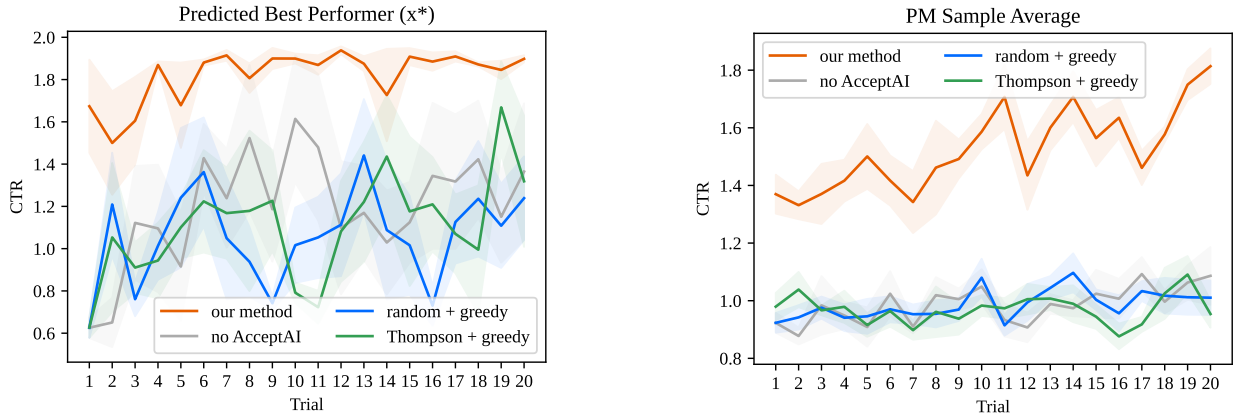
4.5.2 Training Results

Figure 5 displays the trajectory of the peak performance and portfolio quality averaged across the five independent simulation runs. The results demonstrate a rapid convergence capability. As seen in Panel (a), our method (orange line) identifies the region with the global acceptable optimum ($CTR = 2\%$) of the performance landscape within five to seven trials.

Beyond identifying a single optimal point, the framework demonstrates robust learning of the high-performing region. The portfolio quality metric in Panel (b), which measures the average performance of a batch drawn from the model's posterior distribution of possible maximizer locations, tracks closely with the peak performance. This indicates that the model

concentrates its probability mass around true optima, allowing for the consistent generation of high-quality creative options rather than relying on a lucky outlier.

Figure 5: Performance Comparison of the Active Learning Framework against Benchmarks.



(a) Peak Performance (x^*_{pred})

(b) Probability Matching Sample Average

Notes: Comparison of the proposed Active Learning framework (Orange) against Random (Blue), Thompson Sampling (Green), and No-AcceptAI Ablation (Grey). Shaded areas represent the standard error over 5 independent runs.

Comparing these results with the benchmarks confirms the superiority of our proposed framework. As evident in Figure 5, our method (orange) rapidly dominates all benchmarks, which show only marginal improvement over time and fail to reach the performance levels of our approach. This confirms that the 304-dimensional space is too vast for other exploration strategies to succeed. Benchmarks suffer from a "local optima trap," frequently stalling at the 1% CTR regions rather than effectively exploring towards the 2% global maximum. The failure of Thompson Sampling (TS) highlights a critical limitation: TS is designed to maximize the number of clicks given a budget (optimizing exploitation), not to find the top creative given a budget (pure exploration). The space being high dimensional, the uncertainty remains high on the vast majority of the space at all time. If the expected reward in this high-uncertainty region is lower than the current known local optimum, TS will allocate very little resources to exploring this space. It is thus naturally biased toward exploiting regions of known moderate success: it prioritizes refining its knowledge of these "good enough" regions rather than risking resources on the aggressive exploration required to find rare peaks in high-dimensional space.

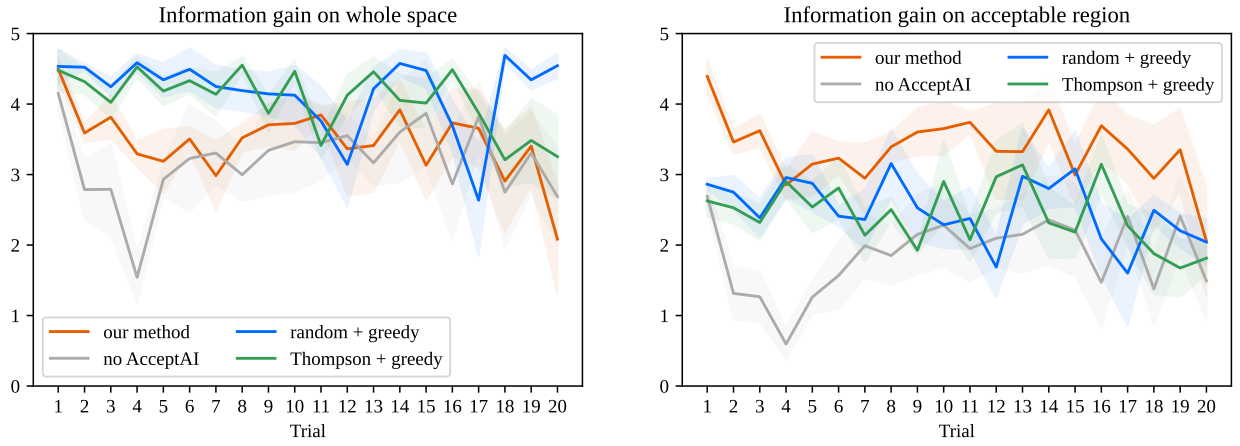
4.5.3 Information Gain Analysis

To understand why our method outperforms the benchmarks, we analyze the *Realized Information Gain* to see how much each batch reduces the model's uncertainty about the location of an optimal creative. A higher information gain implies the system is actively "learning" where the best designs are located.

However, not all information is equal for our purpose. Figure 6 reveals a critical distinction. Panel (a) shows the information gain calculated over the *entire* creative space. Here, the benchmarks (Random and Thompson Sampling) perform well, often exceeding our method. This is because they are efficiently learning "negative knowledge', confirming that vast regions of the 304-dimensional space are low performing. While this is very informative and reduces global entropy, it does not necessarily help locate the specific acceptable high-performance region we care about.

In contrast, Panel (b) presents the information gain calculated specifically over the *acceptable subspace*. We compute this by restricting the domain of the posterior distribution to the ground-truth acceptable regions (where $S_A(x) \geq 0.5$). Here, our method (orange line) dominates. By explicitly weighting the acquisition function by the probability of acceptability, our system ignores the irrelevant dark space and focuses its "learning budget" entirely on resolving the shape of the performance peaks within the acceptable domain.

Figure 6: Realized Information Gain across the search process.



(a) Gain on Whole Space

(b) Gain on Acceptable Space

Notes: While benchmarks reduce global uncertainty effectively (a), our method is more efficient at reducing uncertainty within the relevant acceptable regions (b).

4.5.4 Efficiency of Pre-filtering for Acceptability

The ablation study (grey line in Figure 5) underscores the necessity of the AcceptAI filter. Without it, the unconstrained model attempts to learn the landscape of the entire 304-dimensional space. While it eventually locates high-performing regions, around 80% of the sampled points are located outside the acceptable domain. While we provided the CTR results for these unacceptable points (overriding the vetting process), this resulted in fewer valid data points per batch allocated to the relevant regions of the space compared to the full framework. This confirms that pre-filtering the search space using a dedicated brand-alignment model is a crucial component for practical, cost-effective creative optimization.

5 Field Experiment

5.1 Objectives and Experimental Setting

Following the successful validation in a controlled simulation, we deploy our framework in a live field experiment to evaluate its end-to-end performance in a real-world setting. The primary objective is to demonstrate that our system can efficiently learn a firm's subjective brand standards and objective performance criteria from a limited number of costly field deployments, ultimately generating creatives whose performance is competitive with managerially relevant benchmarks.

The experiment is conducted with a partnering company seeking to optimize background images for its outdoor activities advertising campaign on Instagram. All campaigns target an audience of residents aged 18 to 55 who are parents of children aged three to twelve. Placements are manually restricted to Instagram's feed, Explore, and search results. The training phase takes place in Q4 2023, a period of lower stakes for the company (winter season). The primary performance metric is the link click-through rate (CTR). Each ad creative consists of a standardized template containing a logo and text overlay, as illustrated in Figure 7, ensuring that only the background image varied across conditions. The template was created by a company-approved designer with over 5 years of experience in the field.

Figure 7: Simplified representation of the template provided by the company.

5.2 Navigating the Live Advertising Ecosystem

A central challenge in live field experiments is isolating the creative's effect from the complex dynamics of the ad platform. Modern advertising platforms like Instagram use a secondary optimization algorithm to manage ad delivery, meaning a creative's performance is jointly determined by its intrinsic appeal and the platform's allocation decisions. In the case of experiments studying causality, this might introduce concerns regarding path dependence and algorithmic bias.

We treat this algorithmic behavior as an inherent feature of the ecosystem rather than a confounding factor. Our experiment is not designed to establish a pure causal effect of a visual element on a specific consumer, but rather to predict the outcome of an ad's deployment within the existing delivery system. In this view, the platform acts as a stochastic

response function. Our goal is to find creatives that are effective *within* this real-world environment. To quantify the stability of the performance signal, we conducted a reliability test by deploying the same batches on two subsequent periods, yielding an average Spearman's correlation of 0.43. While this confirms the presence of noise from the platform's optimization and other temporal factors, this correlation indicates that a substantial and learnable performance signal is attributable to the static quality of the background images themselves.

5.3 Phase 1: Learning Brand Standards

The first phase focuses on training the AcceptAI model to understand the partner firm's brand standards, ensuring that brand-safe visuals are selected for the costly field tests. For this application, "acceptability" is defined as visuals that are (1) photorealistic, (2) representative of natural landscapes in the company's service areas, and (3) free of AI-generated artifacts.

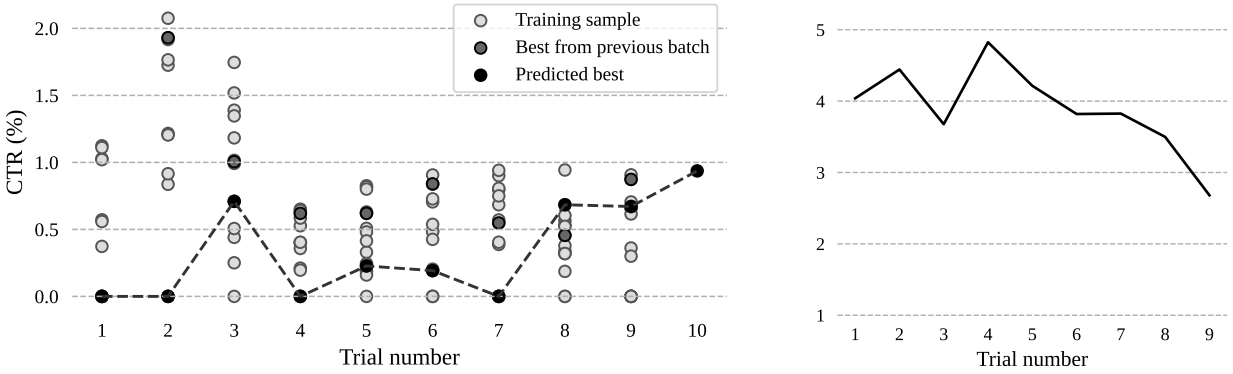
Unlike in the simulation where the raw acceptability score $S_A(x)$ was provided to the active learning engine, the discretized scale introduced in Section 3.5 is used for the field experiment: unacceptable (10%), barely acceptable (50%), good (70%), and great (90%). The model is then trained over 60 batches of 12 visuals. This sequential process was highly efficient, requiring a total of less than two hours of manual labeling time from the company representative to train the model.

5.4 Phase 2: Learning Performance with PerfAI

With the pre-trained AcceptAI model serving as a filter, we train the PerfAI model over 9 trials. In each trial, the active learning engine proposes an informative batch of 12 creatives, which are run alongside the creative predicted to perform best by the model and a seasonal control. From trial 2 onwards, a second control creative is added as the best performer from the previous trial (seasonal control creative), to allow the model to de-trend temporal effects and better estimate the relationship between visual embeddings and performance.

The vetting process for these proposed batches serves the dual purpose of ensuring campaign safety and further refining the brand-alignment model. In our application, this iterative vetting results in an additional 39 batches containing rejected images that are used to update the AcceptAI model. The successfully vetted batch is then deployed to collect CTR data, which is used to update the PerfAI model. Ultimately, the AcceptAI model is trained with 1,417 images and the PerfAI model with 125 images. Both numbers remain excessively low by machine learning standards to train a model with 304 predictors, highlighting the data efficiency inherent in the Bayesian active learning framework.

Figure 8 illustrates the learning progression across the nine training trials. It is important to note that these batches were designed to be maximally informative for training the model, not to maximize immediate CTR. Consequently, the results show a pattern consistent with exploration: the results are highly variable with each batch, while the information gain remains high (Figure 8b). The performance of the predicted best creative is underwhelming in the first training batches (solid black dots in Panel a), a result probably stemming from model overfitting and overconfidence. From trial 8 onwards, the performance becomes both higher and more stable, indicating that the system successfully learned and converged on



(a) Observed CTR (Training)

(b) Realized Information Gain

Figure 8: Learning progression during Phase 2. Panel (a) illustrates the CTR of training batches, where performance stabilizes from trial 8 onwards. Panel (b) shows the reduction in entropy over trials.

a promising region of high-performing visuals. We also note that a major national holiday resulting in a long weekend falls during the testing of the second batch, explaining the unusually high CTRs observed that week.

Further analysis reveals an intelligent search pattern. As visualized in Online Appendix D, Figure 11, the algorithm learned to concentrate its exploration on a specific region of promising *layouts* while still sampling a broad range of stylistic variations (e.g., weather, lighting). This demonstrates a sophisticated learning process that goes beyond simple convergence, identifying a promising creative territory rather than just a single optimal point.

5.5 Final Validation: Head-to-Head Benchmark Comparison

After the 9-week training period, we conducted a final validation test to compare our fully trained system against robust benchmarks.

5.5.1 Validation Benchmarks and Rationale

To provide a comprehensive assessment, we first generated a batch of ten creatives using our trained model. For this, we employed the same probability matching mechanism as in the Simulation, instead of greedily selecting the single visual with the highest predicted CTR. This is managerially more valuable than a single-point prediction, as it provides a set of viable creative options for a campaign.

We compared this batch against two relevant alternative strategies. The first benchmark represents the *human designer* status quo. The same company-approved graphic designer who produced the templates was tasked with creating a batch using their own ten background images. This provides a direct comparison of our system's output against skilled human intuition and professional experience, which is the default approach for most firms.

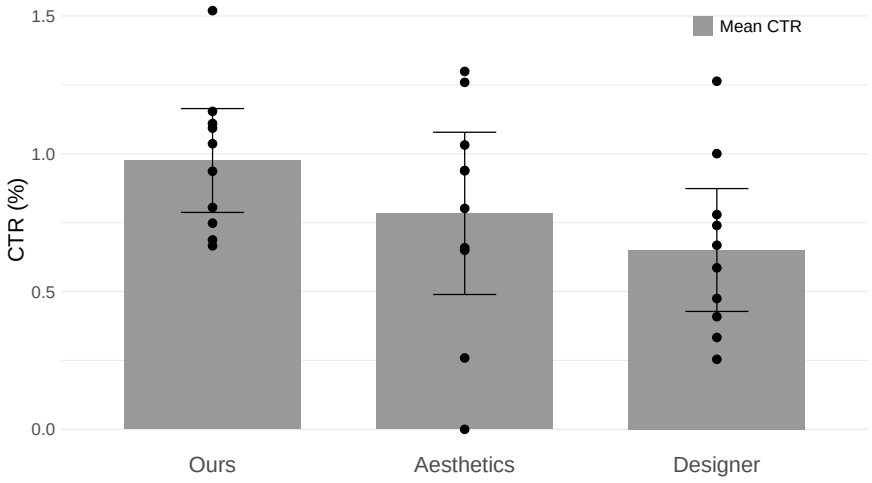
The second benchmark, the Aesthetics Model, serves a dual role as both a conceptual control and a representative of unguided AI generation. This benchmark addresses a common

and tempting shortcut in the industry: using pre-trained off-the-shelf AI models to score visuals or generate content based solely on general beauty rather than empirical performance. To implement this approach, using our GenAI pipeline, we generate ten creatives optimized to maximize an aesthetic score produced by TANet (He et al., 2022), a high-performing image aesthetics assessment model available at the time of our experiment. TANet has been shown to correlate strongly with human aesthetic judgments, reporting a Spearman's rank correlation of 75.8% with human raters. By optimizing exclusively for aesthetic quality, this benchmark allows us to test whether general visual appeal serves as a sufficient proxy for advertising performance, or whether campaign-specific performance learning provides incremental value. By using the same underlying generative capacity as our primary model but omitting our sophisticated active learning guidance, this control isolates whether our custom-trained PerfAI has learned a performance signal specific to the advertising context. Ultimately, this benchmarks our system against an unguided application of state-of-the-art generative models and establishes the marginal value of our performance-driven optimization over a baseline of generic high-quality image generation.

5.5.2 Experimental Results

To ensure a rigorous and fair comparison, all creatives from the three conditions were first vetted and deemed acceptable for deployment by the company. This ensured that all tested images met the baseline brand standards, allowing for a direct comparison on the dimension of performance. The three sets of creatives were then run concurrently during a week, with identical budgets and targeting parameters, to eliminate temporal and budgetary confounds.

Figure 9: Final validation performance of our model's selections (left), the aesthetics-optimized sample (middle), and the human designer's selections (right) with their 95% confidence intervals.



The results are presented in Figure 9 and Table 2. Our AI-generated creatives achieved the highest mean CTR, outperforming both the human designer and the aesthetics-optimized benchmarks. We note that the top achieved CTR in each batch are all between 1.26% and

Table 2: Experimental Results (CTR %)

Group	Count	Mean	Std	Worst	2nd Worst	2nd Best	Best
Ours	10	0.98	0.26	0.67	0.69	1.15	1.52
Aesthetics	10	0.78	0.41	0.00	0.26	1.26	1.30
Designer	10	0.65	0.31	0.25	0.33	1.00	1.26

1.52%. These CTR scores only consider link clicks, which is the metric used to train our model. If clicks on the Instagram posts are also counted, the CTRs become 1.63% (ours), 1.50% (aesthetic), and 1.42% (designer), which is in line with average CTRs reported in that geographical market[2]. Crucially, our approach also exhibited the lowest variance in performance, indicating an ability to *reliably* generate high-performing content and reduce the creative risk inherent in advertising campaigns.

To formally assess the differences between these groups, we conduct a pairwise bootstrap analysis (10,000 iterations). The results indicate that our system's selections outperformed the human designer's batch in 99.59% of samples ($p = 0.0041$). When compared to the aesthetics-optimized benchmark, our system produced a higher mean CTR in 90.76% of the bootstrap samples ($p = 0.0924$). Finally, the aesthetics benchmark outperformed the human designer in 80.29% of the samples ($p = 0.1971$).

While the small sample sizes and the reliance on a single human benchmark warrant a cautious interpretation, these results suggest that our system's performance is at least "acceptable" and competitive by professional standards. Directionally, the results indicate that our system not only learned a performance signal superior to a generic aesthetics model but also produced a portfolio of creatives that were more consistently effective (lower standard deviation) than those selected by the human designer in this instance. These findings highlight the potential for campaign-specific optimization to uncover high-performing visual configurations that may remain overlooked by both human intuition and general aesthetic heuristics.

5.5.3 Long-term Performance

To evaluate the durability of the learned visual features, we conducted a second validation test eighteen months after the initial training phase (April 2024). In a typical managerial implementation, the PerfAI model would be updated continuously using a rolling window of new field data to adapt to environmental changes. However, due to logistical constraints and the significant costs associated with retraining for a new seasonal cycle, the selected creatives from the original winter campaign were deployed without further updates. While not the ideal operational setup, this allows us to evaluate whether the high-performance regions of the creative space identified in the first campaign remained relevant across a significant temporal gap and a distinct seasonal shift (Spring vs. Autumn/Winter).

This retest occurred during a "high-stakes" period, the peak booking season for spring holidays. We compared the top 10 creatives identified by our original model against the "best-

[2]https://www.xyzlab.com/meta-ads-benchmarks/japan reports CTR between 1.3% to 1.9% (Oct 2024).

effort" batch produced by the company's internal team for their actual campaign. Both sets ran concurrently with identical targeting and budget parameters. Table 3 summarizes the results.

Table 3: April High-Stakes Benchmark Results (CTR %)

Group	Count	Mean	Std	Worst	2nd Worst	2nd Best	Best
Ours	10	3.38	0.76	1.86	2.53	3.95	4.32
Company	10	3.24	0.83	2.51	2.55	3.58	5.39

The results demonstrate remarkable stability in the learned performance signals. Even without retraining to account for seasonal drift, our batch achieved a mean CTR (3.38%) slightly higher than the company's campaign (3.24%), with a marginally lower standard deviation (0.76 vs 0.83).

To formally evaluate these results, we conducted a bootstrap analysis (10,000 iterations) to compare the mean performance of the two groups. Our batch achieved a higher mean CTR than the company's contemporary campaign in 67.7% of the bootstrap samples (one-tailed $p = 0.323$). The company's performance was significantly bolstered by a single home-run creative (5.39% CTR), which was a clear outlier compared to their second-best ad (3.58%). Conversely, our system's mean was dragged down by a single "strikeout" (1.86% CTR), while our second-lowest ad (2.53%) was comparable to the company's low performers. When moving past these single-ad anomalies, the AI-generated portfolio shows a more concentrated mass of high-performing ads. Notably, the next five best-performing ads across both groups were generated by our system. These findings suggest that the system successfully captured durable visual attributes that continue to elicit consumer response regardless of seasonality.

We interpret these results as a conservative lower bound of the system's potential. While the company's best individual ad outperformed our best ad, likely due to the human designer's ability to hand-pick historically "safe" winners, our portfolio-wide performance remained consistent. In practice, a firm would likely achieve even greater gains by retraining the model on the higher CTR data obtained close to such a high-stakes period. This suggests that while human designers may still excel at identifying individual high-performing outliers, our framework offers a robust alternative for maintaining consistent portfolio-wide performance with reduced creative risk.

5.6 Field Experiment Findings and Additional Insight

The field experiment successfully demonstrates the practical viability of our framework. The results show that the system can efficiently learn both subjective brand constraints and objective market performance within a complex, real-world advertising environment. The final head-to-head comparison provides strong evidence that our automated approach can generate a portfolio of creatives that are not only competitive with traditional human-led processes, but also consistent.

A key insight from our trained models is that the set of visuals predicted to be top performers is largely distinct from the set deemed highly brand-acceptable (see Online Appendix

D, Figure 13). We found that only 5.3% of visuals in our extensive training set fell into the top quartile for both metrics. When retaining only top deciles, the overlap goes down to 0.6%. This highlights the critical "needle in a haystack" challenge for managers: the creatives most likely to drive clicks are often not the ones that align with the brand's core identity or campaign requirements. This finding justifies the need for a dual-objective optimization system like ours, as a simple search for high performance would likely yield off-brand and unusable assets.

Figure 10: The final set of ten diverse creatives selected by our system via probability matching for the validation test.

The final set of ten creatives selected by our system (Figure 10) was not monolithic, containing a diverse range of scenes including sea shores, hills, and waterfalls. This illustrates the framework's ability to provide a manager with a portfolio of high-quality, varied options, which could be more valuable than finding a single "best" image, which could be expected with a greedy optimization algorithm. This provides a robust proof-of-concept for the value of integrating generative AI with our active learning methodology for digital advertising.

Finally, we want to discuss the economic and computational accessibility of our framework. The active learning update step, which includes retraining the Bayesian neural networks and optimizing the next batch, requires on average 15 minutes on a standard laptop GPU (RTX 4080 Model). This low computational overhead ensures that the optimization loop can be executed rapidly without requiring enterprise-level infrastructure.

Regarding data acquisition, the primary financial barrier is the cost of media on the advertising platform. In our field experiment, the average Cost Per Mille (CPM) was approximately $7. Thus, obtaining performance data for a batch of 12 creatives with 4,000 impressions each requires an ad spend of roughly $340. A complete nine-trial optimization cycle, matching the scope of this study, requires a total media investment of approximately $3,000 plus modest computational expenses.

By comparison, outsourcing equivalent work to a creative agency typically involves significantly higher costs, although a direct comparison remains complex as fees vary based on professional reputation, the scale of the application, and the potential necessity for custom photography or physical photoshoots. Agency workflows may also follow entirely different

schedules regarding the number of trials and creatives per batch, and they offer no inherent guarantees of finding superior creatives through traditional processes. To provide a representative baseline for a nine-trial cycle, we estimate a total agency cost of approximately $13,000. This includes a conservative $9,000 in creative production fees[3] and approximately $4,025 in management fees (typically 15% of total managed media spend[4]). This comparison illustrates how our framework democratizes access to high-end creative optimization for smaller advertisers by shifting the barrier from expensive labor to manageable media spend.

6 General Discussion and Conclusion

Generative AI is expected to redefine the landscape of digital advertising, shifting the core managerial challenge from the manual production of a few creative assets to the strategic navigation of a vast design space. This shift introduces a dual problem for managers: a *search problem* of identifying high-performing visuals within a high-dimensional latent space, and an *alignment problem* of ensuring those visuals adhere to brand standards. This paper introduces and validates an integrated framework that addresses these challenges. By combining a specialized generative model with a dual-objective Bayesian active learning engine, our system demonstrates that systematic exploration of creative content can be both computationally feasible and data-efficient, even under the high costs of in-field data acquisition.

The primary methodological contribution of this research is a Bayesian active learning framework designed to solve the dual-objective constrained search problems in high-dimensional creative spaces. By leveraging Bayesian Neural Networks as flexible surrogate models, the system provides the uncertainty quantification necessary to guide an efficient information-driven search. This approach helps mitigate the data scarcity inherent in live advertising campaigns, where each observation requires an expensive field experiment. Furthermore, our empirical results from a live advertising ecosystem demonstrate that this system can generate portfolios of creatives that achieve performance and consistency comparable to those of professional human designers and unguided AI models optimized for general aesthetics. This provides a proof-of-concept for the viability of automated data-driven creative optimization.

A critical managerial insight uncovered by this work lies in the fundamental nature of the creative space itself. Our application reveals that there exists in some cases a striking lack of alignment between visual features that drive consumer engagement and those that satisfy brand guidelines. In our training sample, only 5.3% of visuals fall into the top quartile for both metrics, and when the criteria are tightened to the top deciles, this overlap collapses to a mere 0.6%. This "needle in a haystack" reality underscores the value of a dual-objective architecture. It suggests that a search focused solely on performance is likely to yield off-brand assets, while a narrow focus on brand-safe aesthetics may overlook high-performing configurations that resonate with consumers.

These findings have implications for creative strategy and execution. First, our framework suggests a systematic approach to creative testing that moves beyond the traditional "test-

[3] `https://mthdmarketing.com/blog-posts/graphic-design-agency-complete-guide-to-services-pricing-selection`

[4] `https://mthdmarketing.com/blog-posts/marketing-agency-pricing`

and-select" model of A/B testing or multi-armed bandits. By integrating the generation process directly into the learning loop, the approach enables a more structured exploration of the creative continuum, not just selection from a pre-defined set.

Second, this methodology offers a potential path toward democratizing access to high-end creative optimization. While the initial implementation currently requires significant technical expertise to build and maintain the Bayesian pipeline, the framework might also be of interest to advertising platforms or third parties with technical expertise who could provide such optimization as a service. By establishing a cost-effective approach to data acquisition, the framework enables firms with limited media budgets to train predictive models from a small number of field tests, thereby reducing the financial risk associated with large-scale creative testing. As generative tools and machine learning platforms become more standardized, this approach could make data-driven creative development increasingly feasible for a wider range of advertisers.

Third, for larger firms, our framework should be viewed as a tool for augmenting human creativity. The system can generate, test, and evaluate novel visual concepts at scale, providing human creative teams with data-driven insights into which specific layouts and visual elements are most effective for their target audience. This allows designers to focus on higher-level strategy and brand storytelling, supported by a deeper empirical understanding of the creative landscape.

The effectiveness of these creative insights relies on a clear understanding of the framework's current boundary conditions and the necessity of operational rigor. As a proof-of-concept, our system is optimized for "cold-start" scenarios, focusing on background images where the optimization engine can navigate the design space efficiently. While the framework scales with larger batches and more general models, it currently faces limitations in semantic precision, such as exact text rendering or complex object interactions, that may require deeper architectural interventions. Beyond these technical constraints, the framework requires managers to maintain strict experimental control to yield valid results. Targeting and bidding specifications (including audience segments, budgets, and placements) must remain constant across all tested creatives to ensure that observed performance differences are driven by visual content rather than platform-side delivery dynamics. In our field experiment, we strictly controlled these parameters to ensure CTR variances were mostly attributable to the creative variations themselves.

Transitioning from background generation to more intricate creatives, such as those featuring specific products or complex human figures, will require more sophisticated generative architectures. These domains demand high-fidelity semantic control, including exact text rendering, that remains a challenge for current diffusion models. Addressing these systemic limitations likely requires deeper structural interventions beyond the optimization logic of our current engine. Furthermore, future research should aim to enhance operational efficiency by integrating Vision-Language Models (VLMs) to automatically codify brand guidelines for AcceptAI when possible. Such automation would significantly reduce the manual labeling burden and provide more informed priors for scoring models.

Finally, a significant opportunity exists to address the contextual specificity of our performance models. Currently, these models are tied to the specific product, audience, and platform of a single campaign. Investigating transfer learning techniques to determine if visual effectiveness patterns generalize across different contexts would be a valuable extension.

Successfully carrying over performance signals from one campaign to another would mitigate the need for full training cycles for every new ad set, thereby accelerating the deployment of optimized creative content across a firm's entire marketing portfolio.

The proliferation of generative AI marks a significant shift for creative industries. The challenge for firms is evolving from a scarcity of creative content to an abundance of it. Success will increasingly depend on the ability to intelligently navigate these new, vast design spaces to discover content that is both effective and brand-aligned. This paper provides a foundational framework for that navigation, bridging the gap between the theoretical potential of generative AI and its effective application in digital advertising.

A VAE-GAN for Layout Generation

The generative model for semantic layouts is a Variational Auto-Encoder combined with a Generative Adversarial Network (VAE-GAN). This architecture is chosen for its ability to learn a smooth and structured latent space while producing high-quality, realistic outputs. The model consists of three main components: an Encoder, a Decoder (or Generator), and a Discriminator.

A.1 Model Architecture

The Encoder network is designed to map an input semantic layout image, L, to a distribution in the 48-dimensional latent space. Specifically, it outputs the parameters of a diagonal Gaussian distribution: a mean vector $\mu \in \mathbb{R}^{48}$ and a log-variance vector $\log(\sigma^2) \in \mathbb{R}^{48}$. A latent vector x_L is then sampled from this distribution using the reparameterization trick, $x_L = \mu + \sigma \odot \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. The encoder is composed of a series of convolutional layers with max-pooling and batch normalization, which progressively downsample the input to produce the latent parameters.

The Decoder (Generator), G, maps a point x_L sampled from the latent space back to the image space, producing a generated layout $\hat{L} = G(x_L)$. Its architecture is symmetric to the Encoder, consisting of a series of transposed convolutional layers with upsampling and batch normalization to reconstruct the layout's original dimensions.

The Discriminator, D, is a convolutional neural network trained to distinguish between real layout images from the training data and fake layouts generated by the Decoder. It takes a layout image as input and outputs a scalar probability indicating whether the image is real or generated.

A.2 Training Data and Pre-processing

The VAE-GAN is trained on a dataset of semantic layouts extracted from a large corpus of landscape images. The image corpus contains 135,719 pictures, combining 45,719 images tagged as "landscape" from the Flickr API and 90,000 images from the Landscapes High-Quality (LHQ) dataset (Skorokhodov et al., 2021).

For each image, we perform a pre-processing step to obtain a semantic layout. We use a pre-trained DeepLabV2 model (Chen et al., 2017) to perform semantic segmentation, which assigns a categorical label (e.g., sky, tree, water, mountain) to each pixel. The resulting segmentation maps serve as the training data for our VAE-GAN.

A.3 Training Objective

The model's parameters are trained jointly by optimizing a composite loss function that combines objectives from both the VAE and GAN frameworks. The overall objective is to learn a latent space that allows for both accurate reconstruction and the generation of realistic, novel layouts. The loss function consists of three primary components:

1. **The VAE Reconstruction Loss ($\mathcal{L}_{\text{recon}}$):** This term encourages the Decoder to produce outputs that are faithful to the original inputs when passed through the Encoder. It is measured as the pixel-wise cross-entropy between the input layout L and the reconstructed layout $\hat{L}$.

2. **The Latent Loss ($\mathcal{L}_{\text{latent}}$):** This is the Kullback-Leibler (KL) divergence between the Encoder's output distribution, $q(x_L|L)$, and a standard normal prior, $p(x_L) = \mathcal{N}(0, I)$. This term acts as a regularizer, encouraging the Encoder to learn a smooth and well-structured latent space.

$$\mathcal{L}_{\text{latent}} = D_{KL}(q(x_L|L)||p(x_L)) \tag{12}$$

3. **The Adversarial Loss ($\mathcal{L}_{\text{adv}}$):** This loss is derived from the GAN framework. The Decoder (Generator) G is trained to minimize this loss by producing layouts that the Discriminator D misclassifies as real. The Discriminator is simultaneously trained to maximize it by correctly distinguishing between real and generated layouts. This adversarial dynamic pushes the Decoder to generate increasingly realistic images.

$$\mathcal{L}_{\text{adv}} = \min_G \max_D \mathbb{E}_{L\sim p_{\text{data}}(L)}[\log D(L)] + \mathbb{E}_{x_L\sim p(x_L)}[\log(1 - D(G(x_L)))] \tag{13}$$

In practice the expectations are approximated by averaging over batches of data samples. The complete training objective combines these losses to train the Encoder, Decoder, and Discriminator networks simultaneously, ensuring that the model learns to generate a diverse and realistic set of landscape layouts from the learned 48-dimensional creative space.

B Bayesian Neural Network Architecture

The AcceptAI and PerfAI models are both Bayesian neural networks (BNNs) and share the same architecture. This architecture was determined with the simulation exercise and designed to identify a structure that was sufficiently expressive to capture the complex relationships in the creative space while also converging rapidly during training.

The retained model is a fully connected feed-forward neural network. The input layer has 304 nodes, corresponding to the concatenated 48-dimensional layout embedding and the 256-dimensional style embedding. The network has four hidden layers with sizes [128, 64, 32, 8] and a single output node. All hidden layers use a Leaky ReLU activation function, while the final output layer uses a Sigmoid activation function to scale the prediction to a probability for AcceptAI or to a bounded CTR within the range [0-10%] for PerfAI.

To account for the seasonal effects in the PerfAI model, which arise from testing batches sequentially over time, we augment the network with a set of control parameters θ_S. For each batch t, we provide an indicator vector z_t where the current period is indicated by a 1 (and the rest is 0) as an additional input to the model's forward pass. The θ_S parameters are multiplied by the indicator vector and added to the output of the final layer just before the sigmoid activation function. The final output for an embedding w in batch t is therefore computed as

$$\text{sigmoid}(f_{\text{NN}}(x; \theta_{NN}) + z_t^\top \theta_S). \tag{14}$$

This allows the model to learn a separate intercept for each batch, effectively capturing time-based shocks to the baseline CTR without confounding the learned relationship between the visual embeddings and performance.

C Computational Implementation

C.1 Hamiltonian Sequential Monte Carlo

Due to the multi-modal nature of the posterior distribution generated by neural networks, conventional MCMC methods like Metropolis-Hastings or standard Hamiltonian Monte-Carlo (HMC) can struggle to explore the parameter space efficiently. To overcome this, we employ Hamiltonian Sequential Monte Carlo (HSMC), which combines sequential Monte Carlo (particle filtering) with HMC-based mutation steps (Burda and Daviet, 2022). This method is particularly adept at sampling from high-dimensional, multi-modal posteriors.

The HSMC algorithm maintains a set of H weighted samples, or "particles," $\{\theta_h, \omega_h\}_{h=1}^{H}$, which represent the posterior distribution $p(\theta|\mathcal{D})$. When new data arrives, the algorithm proceeds in three steps:

1. **Correction:** The weight ω_h of each particle θ_h is updated based on the likelihood of the new data under that particle. This is an application of importance sampling, where particles that better explain the new data are assigned higher weights.

2. **Selection:** A new set of particles is resampled from the old set based on their updated weights. Particles with higher weights are more likely to be duplicated, while those with low weights are eliminated. This step focuses computational effort on more promising regions of the parameter space.

3. **Mutation:** Each resampled particle is evolved using several steps of an HMC transition kernel. This gradient-informed step allows the particles to efficiently explore the local geometry of the posterior distribution, leading to a new set of diverse and high-probability samples.

This sequential process allows the set of particles to adaptively track the posterior distribution as more data is incorporated. In our application, we set $H = 128$ and perform 16 mutation steps at each iteration.

C.2 Computing Information Gain

The core challenge is to efficiently approximate the KL divergence between the prior, $p(\mathcal{X}|\mathcal{D}_{\text{old}})$, and posterior, $p(\mathcal{X}|\mathcal{D}_{\text{new}})$, distributions over the optimal sets. The KL divergence is defined as an integral over the posterior distribution, which is itself intractable.

To solve this, we turn to importance sampling, a Monte Carlo technique for approximating expectations with respect to a target distribution (our posterior) using samples drawn from a different proposal distribution (our prior). In general, to estimate the expectation of

a function $f(x)$ under a target distribution $p(x)$ using samples $\{x_h\}_{h=1}^H$ from a proposal distribution $q(x)$, the formula is:

$$\mathbb{E}_{p(x)}[f(x)] = \int f(x)p(x)dx \tag{15}$$

$$\approx \frac{1}{H}\sum_{h=1}^{H} f(x_h)\frac{p(x_h)}{q(x_h)} \tag{16}$$

where the ratio $w_h = p(x_h)/q(x_h)$ is called the importance weight.

Our application of this principle begins by generating samples from the proposal distribution, which in our case is the prior over optimal creatives, $p(\mathcal{X}|\mathcal{D}_{\text{old}})$. To do this, we find the optimal embedding vector x_h^* for each of the H particles $\{\theta_h\}_{h=1}^H$ from our HSMC sampler by running a gradient ascent optimization for each particle. This resulting collection of optimal points, $\{x_h^*\}_{h=1}^H$, represents our discrete sample. For computational purposes, we then create a smooth density approximation from these samples, $\hat{p}(\mathcal{X}|\mathcal{D}_{\text{old}})$, using a Gaussian Kernel Density Estimator (KDE).

Next, we must approximate the importance weights, which requires the posterior density, $p(\mathcal{X}|\mathcal{D}_{\text{new}})$. We approximate this density using the same weights as those generated during a HSMC correction step. Instead of re-running the expensive optimization for every particle to obtain a new set of samples from the posterior, we can approximate the posterior density $\hat{p}(\mathcal{X}|\mathcal{D}_{\text{new}})$ by constructing a *weighted* Gaussian KDE over our existing set of optimal points $\{x_h^*\}_{h=1}^H$. Each point x_h^* is weighted by the $\omega_h \propto p(\mathcal{D}_{\text{new}}|\theta_h)$ of its corresponding particle.

With these two density approximations, the KL divergence can be estimated using the importance sampling framework, where our function is $f(\mathcal{X}) = \log \frac{p(\mathcal{X}|\mathcal{D}_{\text{new}})}{p(\mathcal{X}|\mathcal{D}_{\text{old}})}$. This gives the estimator:

$$D_{KL}\left(p(\mathcal{X}|\mathcal{D}_{\text{new}})\|p(\mathcal{X}|\mathcal{D}_{\text{old}})\right) \tag{17}$$

$$= \mathbb{E}_{p(\mathcal{X}|\mathcal{D}_{\text{new}})}\left[\log \frac{p(\mathcal{X}|\mathcal{D}_{\text{new}})}{p(\mathcal{X}|\mathcal{D}_{\text{old}})}\right] \tag{18}$$

$$\approx \frac{1}{H}\sum_{h} \frac{\hat{p}(x_h^* \in \mathcal{X}|\mathcal{D}_{\text{new}})}{\hat{p}(x_h^* \in \mathcal{X}|\mathcal{D}_{\text{old}})} \log \frac{\hat{p}(x_h^* \in \mathcal{X}|\mathcal{D}_{\text{new}})}{\hat{p}(x_h^* \in \mathcal{X}|\mathcal{D}_{\text{old}})}, \tag{19}$$

where the densities are evaluated using the unweighted KDE (for the prior) and the weighted KDE (for the posterior) at each sample point x_h^*. This integration approach is numerically stable as it is evaluated at points where the respective densities are not close to zero, avoiding instabilities from the log and division operations.

C.3 Approximating Expected Information Gain

The objective function for selecting the next batch requires calculating the *expected* information gain, which involves an intractable integral over all possible future data outcomes for a candidate batch X^B. We approximate this expectation using Monte Carlo simulation. The simulation proceeds by generating a set of H plausible future datasets, where each plausible future is generated from the predictive distribution of one of the H particles from our HSMC sampler.

For the PerfAI model, generating a plausible outcome for a given particle θ_h involves first predicting the expected CTRs $\mu_h = f_{\text{PerfAI}}(X^B; \theta_h)$. We then obtain the corresponding number of clicks y_h by using the rounded expected value (for 4000 impressions), which keeps the simulation deterministic given the particle. A similar process is used for AcceptAI, where predicted probabilities are converted into expected acceptability labels, interpreted as the expected number of times it would be labeled as acceptable when presented to a set of managers. This process results in a set of H distinct, plausible future datasets, $\{y_h\}_{h=1}^H$.

The next step is to calculate the information gain that would be achieved for each of these simulated futures. For a given plausible outcome y_h, we can calculate the corresponding KL divergence, $D_{KL}(y_h)$, by performing a "what-if" analysis: if we were to observe this outcome, the importance weight for every other particle $\theta_{h'}$ would be updated to $\omega_{h'}(y_h) \propto p(y_h|\theta_{h'})$. These updated weights would then be used to compute a new posterior distribution over the optimal set, $\hat{p}(\mathcal{X}|y_h)$, allowing for the calculation of the resulting KL divergence.

Finally, the expected information gain is approximated by taking the average of the KL divergences calculated across all simulated future outcomes:

$$\mathbb{E}[D_{KL}] \approx \frac{1}{H} \sum_{h=1}^{H} D_{KL}(y_h)$$

This provides a computationally feasible estimate of the objective function, which can then be maximized using a standard gradient-based optimizer to find the most informative batch of creatives to test next.

C.4 Optimization Approach

To find the optimal batch of creatives X^B that maximizes the expected information gain, we need an efficient optimization algorithm. The creative space of embeddings, w, is continuous. Critically, because our entire computational pipeline—from the BNNs to the KDEs used to approximate the information gain—is composed of differentiable functions, the expected information gain objective function $G(X^B)$ is differentiable with respect to the embedding vectors in the batch X^B. This allows us to use gradient-based optimization methods.

We employ gradient ascent, a standard iterative optimization algorithm. The core principle of gradient ascent is to iteratively take small steps in the direction of the steepest increase of the objective function. The direction of steepest ascent is given by the gradient of the function, denoted ∇G.

The process starts with a randomly initialized batch of creatives, X_0^B. At each iteration k, the batch is updated according to the following rule:

$$X_{k+1}^B = X_k^B + \alpha \nabla G(X_k^B)$$

Here, α is the learning rate, a small positive scalar that controls the size of each step. The gradient $\nabla G(X_k^B)$ is a matrix containing the partial derivatives of the objective function with respect to each element of each embedding vector in the batch. When optimizing for a full batch, this differentiation is applied simultaneously to all creatives in the batch. This process is repeated until the value of the objective function no longer improves, at which point the algorithm has converged to a local maximum, yielding the next batch of creatives to be tested.

D Visualization of the Creative Space

Figure 11: UMAP representation of embedding vectors for the training data (dots) and tested visuals from each trial (squares, colored by trial number).

Figure 12: Generated pictures and their position in the UMAP spaces.

(a) Layout Space

(b) Style Space

To provide intuition for the high-dimensional search process, we use Parametric UMAP to project the 304-dimensional embedding space into a 2D visualization. We create separate

projections for the 48-dimensional layout embeddings and the 256-dimensional style embeddings, as this better reveals the distinct search patterns for scene structure versus aesthetic qualities.

Figure 11 shows this UMAP space and the progression of the tested visuals over the nine training trials. Figure 12 shows examples of generated landscapes at the corresponding locations of the UMAP space. The algorithm's search for promising layouts converges to a specific region of the space, while the search for styles remains more exploratory, covering a wider area.

Figure 13: Predicted acceptability scores (top) and performance scores (bottom).

Notes: Points represent top quartiles and crosses indicate the location of the final creatives sampled by probability matching.

Figure 13 visualizes the predicted top quartiles for brand acceptability and CTR performance. We note that the dense regions differ for the 2 criteria, resulting in limited overlap

between these two sets. This underscores the challenge of finding creatives that satisfy both criteria simultaneously.

References

Balandat, M., B. Karrer, D. Jiang, S. Daulton, B. Letham, A. G. Wilson, and E. Bakshy (2020). Botorch: A framework for efficient monte-carlo bayesian optimization. *Advances in neural information processing systems 33*, 21524–21538.

Brochu, E., V. M. Cora, and N. De Freitas (2010). A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*.

Burda, M. and R. Daviet (2022). Hamiltonian sequential monte carlo with application to consumer choice behavior. *Econometric Reviews 618*, 21.

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

Chaloner, K. and I. Verdinelli (1995). Bayesian experimental design: A review. *Statistical science*, 273–304.

Chen, L.-C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence 40*(4), 834–848.

Dew, R. (2024). Adaptive preference measurement with unstructured data. *Management Science*.

Dzyabura, D., S. El Kihal, J. R. Hauser, and M. Ibragimov (2023, November). Leveraging the power of images in managing product return rates. *Mark. Sci. 42*(6), 1125–1142.

Dzyabura, D. and J. R. Hauser (2011). Active machine learning for consideration heuristics. *Marketing Science 30*(5), 801–819.

Feit, E. M. and R. Berman (2019). Test & roll: Profit-maximizing a/b tests. *Marketing Science 38*(6), 1038–1058.

He, S., Y. Zhang, R. Xie, D. Jiang, and A. Ming (2022). Rethinking image aesthetics assessment: Models, datasets and benchmarks. In *IJCAI*, pp. 942–948.

Huang, D. and L. Luo (2016). Consumer preference elicitation of complex products using fuzzy support vector machine active learning. *Marketing Science 35*(3), 445–464.

Izmailov, P., S. Vikram, M. D. Hoffman, and A. G. G. Wilson (2021). What are bayesian neural network posteriors really like? In *International conference on machine learning*, pp. 4629–4640. PMLR.

Karras, T., S. Laine, and T. Aila (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410.

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

Park, T., M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu (2019). Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2337–2346.

Peng, T., G. Gui, M. Brucks, D. J. Merlau, G. J. Fan, M. B. Sliman, E. J. Johnson, A. Althenayyan, S. Bellezza, D. Donati, H. Fong, E. Friedman, A. Guevara, M. Hussein, K. Jerath, B. Kogut, A. Kumar, K. Lane, H. Li, V. Morwitz, O. Netzer, P. Perkowski, and O. Toubia (2026). Digital twins as funhouse mirrors: Five key distortions.

Podell, D., Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach (2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Qwen Team (2025). Qwen3 technical report.

Schwartz, E. M., E. T. Bradlow, and P. S. Fader (2017). Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science 36*(4), 500–522.

Shahriari, B., K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas (2015). Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE 104*(1), 148–175.

Sisodia, A., A. Burnap, and V. Kumar (2024). Generative interpretable visual design: Using disentanglement for visual conjoint analysis. *Journal of Marketing Research*, 00222437241276736.

Skorokhodov, I., G. Sotnikov, and M. Elhoseiny (2021). Aligning latent and image spaces to connect the unconnectable. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14144–14153.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika 25*(3/4), 285–294.

Toubia, O., G. Z. Gui, T. Peng, D. J. Merlau, A. Li, and H. Chen (2025). Database report: Twin-2k-500: A data set for building digital twins of over 2,000 people based on their answers to over 500 questions. *Marketing Science 44*(6), 1446–1455.