Leveraging Generative AI to Create Visual Content in Digital Advertising

Remi Daviet*, Wisconsin School of Business, University of Wisconsin–Madison

Yohei Nishimura*, Wisconsin School of Business, University of Wisconsin–Madison

*equal contribution

Abstract

Generative artificial intelligence 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. Moreover, the large space of variations that can be generated makes it challenging to identify content that is both performing well and compatible with a brand's standards or campaign objectives. This paper addresses these concerns by proposing a novel creative design process combining generative AI with two deep Bayesian prediction models. The first model identifies potential high-performance visuals, while the second assesses acceptability by the brand. Both models undergo sequential training using optimized batches of creatives, allowing us to minimize costs and required training set size. We demonstrate the effectiveness of our approach with a field application to scene setting in ads for an outdoor activity company. Our results show that our approach can generate high performing visuals consistently, although with potentially less variety than what a human designer could produce. By providing a framework guiding the integration of generative AI in digital advertising, this paper seeks to bridge the gap between theoretical potential and effective practical applications.

1. Introduction

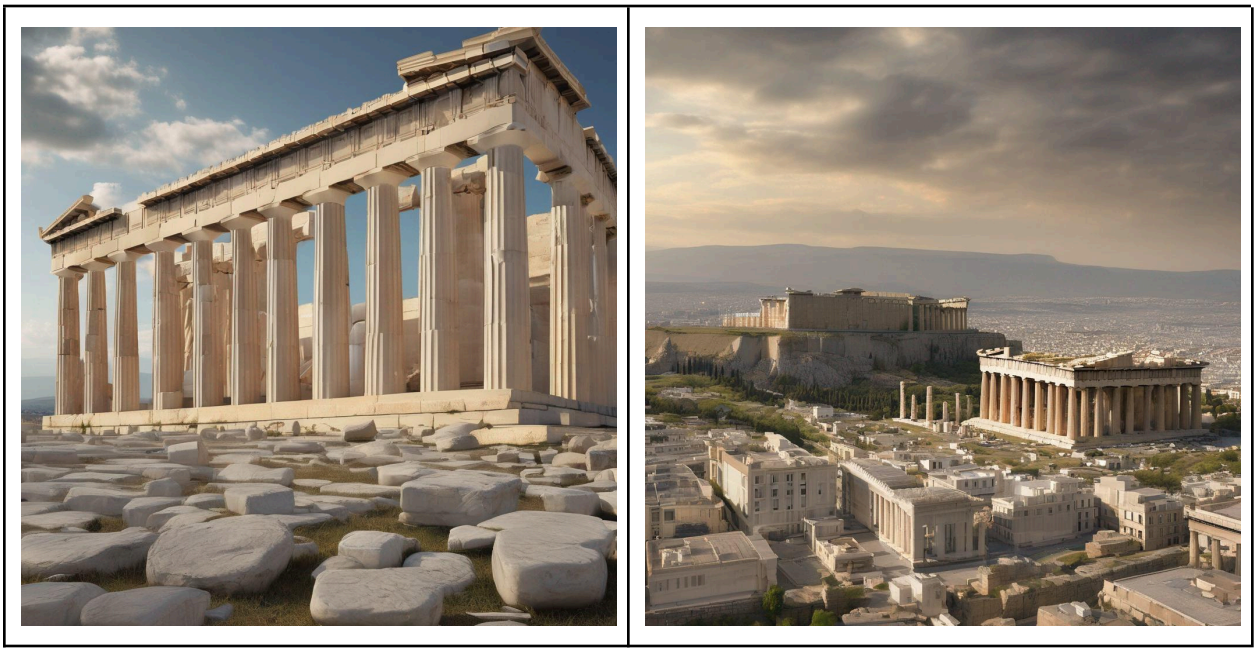
Digital advertising remains one of the dominant approaches to promote products and services (Goldfarb & Tucker, 2011; Johnson et al., 2017; Liu-Thompkins, 2019; Manchanda et al., 2006). An essential element of an online campaign is the process of designing the creatives used. The design decisions might affect a large number of outcomes from advertising effectiveness (Azimi et al., 2012) to brand perception (Gurzki et al., 2019). Traditionally, the production of creatives has been assigned to teams of specialists, from designers to brand managers, bringing their expertise to the collaborative process (de Gregorio & Windels, 2021).

The current processes are subject to high production costs and uncertainty, which can vary depending on the design approach. Designers can for instance use available visuals from large image banks, but selecting images takes time, the images might not possess all the desired characteristics, and digitally editing them is labor intensive. Creative teams can also commission other professionals to produce the desired visuals (e.g., photoshoot, 3D graphics, illustrations…), which consumes a large amount of resources and requires substantial planning. Ultimately, the resulting content heavily depends on both the creative team's expertise and available resources, with no guarantee regarding the advertising performance of each design. Due to these constraints, unless the company has considerable resources, only a handful of creatives are usually produced, and the best performer is selected through experimental methods such as A/B testing or Thompson Sampling (Feit & Berman, 2019; Schwartz et al., 2017).

The emergence of generative artificial intelligence (AI) brings a promising solution to streamline and enhance the process of producing visual content for ad creatives. Firstly, its scalability and efficiency are unmatched as a multitude of design variations can be rapidly generated. This

allows the creative team to substantially reduce the reliance on manual design processes and iterate over variations at a high pace. It could also provide an avenue for exploration of novel creative ideas, leveraging AI to generate designs using new styles and visual elements.

Figure 1. Images generated by SDXL 1.0[1] (Podell et al., 2023) when prompted "A picture of the Parthenon and its surroundings." The images do not accurately represent the Parthenon's environment.



While generative AIs can rapidly generate a large variety of visual content at scale, many unknowns remain regarding their usefulness and performance in a digital advertising context. Modern generative AI architectures work by mapping a numerical vector space—the embedding space—to a space of images corresponding to the training set. The number of dimensions in the embedding space is lower than the number of pixels in an image by orders of magnitudes, usually ranging from 250, for highly specialized AIs, to 2500 for general purpose image synthesis. Intuitively, the embedding vector can be seen as a numerical summary encoding the

[1] SDXL Base 1.0 is the most downloaded image generator on the Hugging face platform, with 4.18 million downloads at the time of writing.

important features of an image. From this summary, the generative model can reconstitute the image, with small variations in an embedding vector leading to small variations in the generated images. Using this approach, a good generative AI can navigate the embedding space to produce an extremely large range of high quality visuals. A currently unsolved problem is the lack of an established method to rapidly identify which parts of the embedding space correspond to visuals that achieve a satisfying performance in an advertising campaign (e.g., high CTR). Moreover, the type of visuals that can be generated depends heavily on the dataset used to train the AI model. A human designer might produce or have access to certain visuals that cannot be produced by a generative AI, as they are either not well represented in the training data (out-of-domain) or they possess specific characteristics that the AI is not capable of reproducing reliably. For instance, while a Generative AI can create beautiful Greek temples, it might not be able to generate an accurate representation of the Parthenon in Athens with its surroundings, even when relevant pictures were in the training sample (see Figure 1). Comparatively, a human designer could easily select such pictures from an image bank. Inaccurate images and representations might be problematic in this context if the goal of the ad campaign is to advertise tourism in Greece. On the other hand, human designers might be limited by their own experience, skills, biases, or risk avoidance, also restricting the space of designs they produce.

Figure 2. Images generated by SDXL 1.0 (Podell et al., 2023) when prompted “People partying” (left) or “Photo of people at a party” (right). There are several issues with the representation of body parts, especially hands.

A second substantial issue is the lack of established methods to ensure that designs generated by an AI satisfy brand standards and requirements. AI has been known to face difficulties with the realism of a number of visual elements (e.g., hands in Figure 2), and with the generation of concepts under-represented in the training sample, leading to either bias or visuals of an unacceptable quality. Beyond the realism and bias issues, a brand might want to have specific designs corresponding with its campaign message, visual identity, or personality. For instance, the brand might want to have recognizable elements that are present in other ads (Schweidel et al., 2006). This leads to questions about the ability to align AI-generated content with what a brand would expect to see in their campaign.

In this paper we propose to address these issues by developing a method which (1) rapidly identifies high performance visuals through sequential batch testing, and (2) maintains the search

for these visuals within the space of designs that are acceptable to the brand. We achieve this by combining several AIs within a Bayesian framework. Specifically, we supplement the generative AI (GenAI) with two predictive Bayesian neural networks: one providing an estimate of the probability that the design will be accepted by the brand (AcceptAI), and one predicting the design's performance (PerfAI). Data sources to train the GenAI are usually available at scale, with many models coming pre-trained on large high-quality datasets. However, training the AcceptAI and PerfAI presents a challenge, as most deep learning predictive models require large datasets, often in the order of thousands of observations, to reach an acceptable predictive accuracy. To train the AcceptAI, generated content needs to be labeled as acceptable or rejected by company professionals whose time is valuable. The PerfAI can be trained with performance metrics obtained through field testing, which becomes rapidly costly and time consuming as the company needs to pay for the test campaigns and wait for the results. Note that even if a visual is predicted to satisfy brand standards by the AcceptAI, a manual vetting is still required before using it in the field, to ensure that no issues are present. This further limits the number of field testable creatives as it is unrealistic to ask brand managers to vet thousands of creatives, regardless of the budget available.

To increase the training efficiency and reduce costs, we propose a Bayesian active-learning approach applicable to both predictive AIs. This allows us to reduce the amount of training data required by an order of magnitude compared to randomly testing images. This approach takes advantage of the fact that we only need accurate predictions for highly acceptable creatives (AcceptAI) and top performers (PerfAI). Indeed, it is irrelevant for us to know with accuracy whether some generated content belongs to the bottom 10% or 20%. Accuracy is only desirable for high scores, which can be rapidly achieved with a carefully designed adaptive training. The

use of Bayesian AIs allows us to adopt a Bayesian sequential optimal design approach, where we generate small batches of highly informative creatives that can rapidly be vetted and tested. The results for each batch are then used to update the AIs and taken into account when producing subsequent batches. For the creatives that are tested in the field, having batches of moderate sizes also allows us to easily vet each of the designs for acceptability, suppressing the risk of undesirable creatives reaching the general public. We show the effectiveness of this approach in a simulation where optimal designs are hidden in a high dimensional space.

We then perform a field test in collaboration with a company promoting activities for families. We generate landscapes used as background for a campaign on nature exploration activities and displayed on a major social media platform. Background images appeared as a natural candidate for our experiment as they have been shown to have an effect on desirable outcomes such as liking and purchase intentions (Maier & Dost, 2018; Yoo & Kim, 2014). We proceed in three steps. We first train our GenAI to produce high-quality landscapes using a combination of publicly available pictures and pre-trained generative models. We then train the AcceptAI to identify generic landscapes corresponding to the geographical locations served by the company and to filter unrealistic visuals (e.g., a tree in the sky). We finally progressively train the PerfAI to predict the Click Through Rate (CTR) by testing in the field a sequence of 9 batches of 14 designs. After training, we generate a batch of 10 creatives predicted to perform well and compare it to two other sets of 10 creatives: one set of generated landscapes that maximize a visual aesthetic score, and one set produced by a professional designer using real pictures. The creatives produced by our approach achieve a mean CTR of 0.98% and a maximum CTR of 1.52%. Comparatively, the generated creatives targeting high aesthetic scores achieved a mean CTR of 0.87% and maximum of 1.30%. Finally, the designer produced creatives reached a mean

CTR of 0.65% and a maximum of 1.26%. Interestingly, the human designer's best two pictures (a starry night and a flower garden) could not have been generated by our AI, as both night pictures and gardens were out of the training domain. Only the designer's fourth best performer (CTR: 0.73%) would fall within the range of styles that our AI could have generated, as it displayed a lake surrounded by mountains.

These results suggest that our approach allows practitioners to train an AI to generate visuals that can achieve a respectable performance, comparable to images selected by a human designer, while maintaining brand requirements and limiting training costs. However, these good results were achieved within a range of visuals that remains limited compared to what can be produced by a human designer. While this paper focuses on generated background visuals for creatives in digital advertising, we believe that the philosophy behind our approach could be adapted to other types of content, including text or audio.

This paper contributes to the field by answering many of the questions surrounding the usability of generative AIs to create visuals in a digital advertising context. First, we show that the recent generation of AIs, capable of producing pictures of sufficient quality to be presented to consumers, can be used with methods affording better control than prompting. Second, we provide a cost-effective approach to maximize the performance of the generated content while maintaining acceptability with regards to brand requirements. Third, we empirically confirm that synthesized content can compete with traditional human-designed content, providing a lower bound for the performance that can be achieved with generative AI. Finally, we discuss some of the limitations currently existing with generated content. We believe that this paper demonstrates the viability of using AIs to generate visuals in digital advertising and leads the way in this rapidly expanding research domain.

2. Business Case

We are working with a company proposing activities for families, such as a day at an aquatic park, attending a cultural festival, or participating in a week-long summer camp including training, exploration, and socializing with others. Customers typically use the platform to search for interesting destinations and to leave reviews of the places they visited. Platform users might have varying goals: some have specific plans (e.g., hiking in a specific area), others can browse a large range of options seeking inspiration. The company is moderately sized with very limited human capital for marketing tasks. Historically, the company has relied on search engine optimization for attracting potential customers and has limited experience in running advertising campaigns. The current objective is to evaluate the potential effectiveness of a campaign on social media, using the promotion of their nature exploration activities as a test case. Provided with a template containing visual elements such as a logo and text, the problem at hand is to identify a background image that would be appealing to potential customers, where the appeal is measured by the click through rate (CTR) of an ad containing the image. The company requires that the background image represents a natural landscape that is both realistic and representative of the locations that the company covers.

One natural option is to hire a human designer to select and edit pictures that they think would be effective. One advantage of this approach is that the designer can easily follow verbal instructions (e.g., “sunny weather”) and likely has relevant experience they can rely on. A constraint is that unless the company has a consequential marketing budget, the number of proposed designs might be limited. The designer might only select designs considered “safe” or

in line with their own experience, as taking creative risks might result in low performing ads and hurt future business prospects.

Another potential option is to use generative AI to produce such backgrounds, given its ability to rapidly provide a large range of visually pleasing landscapes at low cost. The generative AI might be guided to explore potentially risky options as the marginal cost of generating a design is nearly zero. There are however substantial unknowns and limitations with the use of generative technologies. Previous research in marketing has successfully leveraged AI technologies to generate visuals. However, these visuals are usually provided to give insight or guidance to company employees and not directly deployed to face consumers, as their quality is often limited (Sisodia et al., 2023). Moreover, when the space of possible visuals is large, which is the case in our application, generative technologies still require human feedback to be validated or rated (Burnap et al., 2023). In our application, since the generated landscapes are used in ad creatives deployed in the field, they would also need to be vetted by an employee to ensure that the visuals are acceptable. This factor, in addition to potentially prohibitive field testing costs, substantially limits the number of visuals that can be used. Indeed, even with low generating costs, it is unrealistic to expect the company to manually validate and finance the testing of thousands of visuals. Consequently, an efficient method is needed to efficiently train an AI to generate visuals that are both high performing and acceptable from a brand perspective. We propose and test such a method in the following sections.

3. Methodological Approach

Solving the Methodological Problem

Most of the current AI models capable of generating visuals map a lower dimensional space of typically 250 to 2500 dimensions, called the embedding space or latent space, to the space of images. Common architectures include variational auto-encoders (Burnap et al., 2023; Kingma & Welling, 2013), generative adversarial networks (Brock et al., 2018; Goodfellow et al., 2020), or diffusion models which rely on a quantized version of a variational auto-encoders in the background (Podell et al., 2023; Rombach et al., 2021). Provided with a vector *w* representing a location in the embedding space, these models can generate an image or draw from a conditional distribution *p(image|w)*. This embedding space usually possesses some interesting properties and, to some extent, an interpretable structure. First, similar visuals usually correspond to embedding locations that are close to each other. Furthermore, in many cases, points can be interpolated between two embedding locations to progressively change a visual. For instance, points between a location representing a hill and a location representing a mountain should map to progressively higher and higher topographies. As a consequence of the two previous properties, some directions in the embedding space have interpretable meanings: moving in various specific directions can for instance modify the amount of trees, the location of a waterfall, or the type of climate represented. In particular, some methods have been proposed to specifically train a model to provide convenient interpretable embedding space directions (Sisodia et al., 2023). Given these embedding space properties, there are some advantages in directly working in the embedding space for many applications. In our case, searching for high

performance visuals in a structured and lower dimensional embedding space seems more efficient than working in the very high dimensional space of pictures.

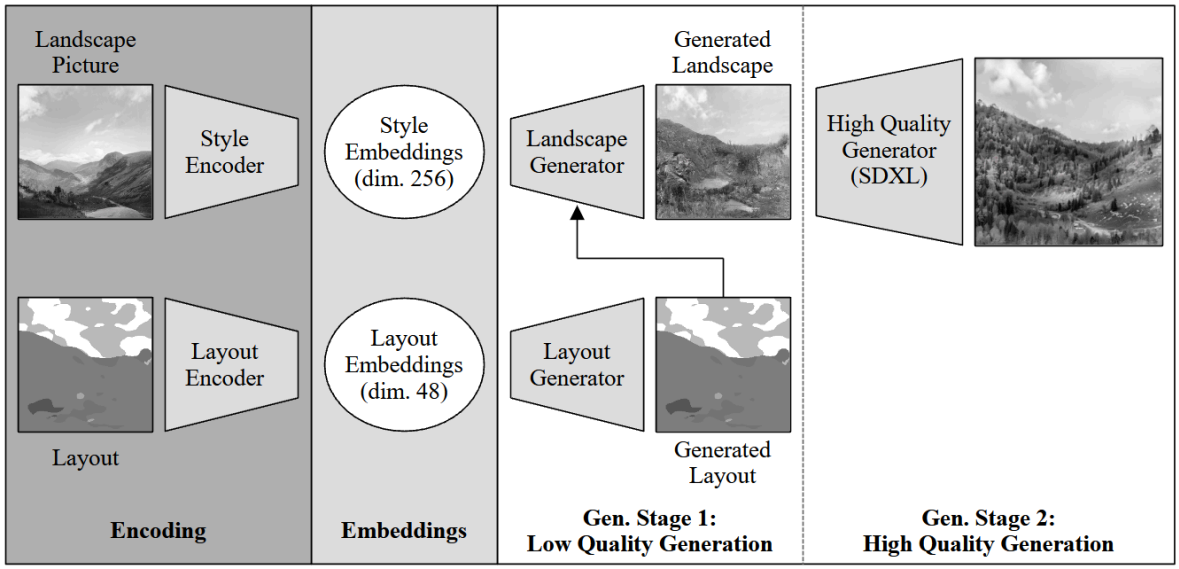
Our first challenge is thus to design and train a generative model that can produce high-quality images and defines that embedding space. We discuss how we solve this problem in the next section. Once an embedding space is defined and pictures can be generated, the second step is to find embedding locations that correspond to pictures that perform well (measured by CTR) and that the company finds acceptable (manually labeled). Ideally, we want to identify various types of landscapes satisfying these conditions, allowing us to maintain some diversity among visuals during the campaign. A simple naive approach would be to randomly search the space and test the generated images to obtain the CTR and the acceptability. Given the dimensionality of the embedding space, an overwhelming amount of data would be needed to find relevant locations. This appears unrealistic, especially for small companies, as manually labeling a large number of images to verify acceptability requires a lot of labor hours and testing ads containing these images in the field to measure CTR is costly. Another approach would be to use traditional optimization methods, but their implementation in our case is not evident. Advanced Bayesian optimization methods have been able to rapidly identify the best options in an embedding space from non-generative models (Dew, 2024). While this is encouraging, as the model was non-generative, the optimization was done only on over a moderate number of existing options (below 1,000). It remains unclear how effective these methods are in a continuous embedding space with unlimited options. Moreover, these methods identify a single optimum, which goes against our need for diversity.

Instead of searching for a single optimal visual, we propose to train two separate AI models that can respectively predict the CTR performance (Perf. AI) and the acceptability (Accept. AI) of a

visual given the corresponding point in the embedding space. Once these models are trained, they can be used to identify several visuals that are both high performing and acceptable. Models predicting an outcome given a generative model's embedding space location have been successfully used in some marketing applications, such as for predicting product characteristics (Sisodia et al., 2023) or aesthetic ratings (Burnap et al., 2023). Note that these papers required large training samples to achieve a degree of acceptable accuracy (respectively 6,187 labeled products and 7,308 ratings). As mentioned before, sample sizes of these magnitudes might be unreasonable, especially if the company is of a moderate size and needs to train these models for a single campaign. To reduce the required training sample size, we adopt a Bayesian approach, which allows us to use Bayesian active learning, the machine learning counterpart to Bayesian experimental design. In the following sections, we formally describe each model, and how they are individually trained.

Generative AI Model

Figure 3. The general architecture of our generative model. Existing styles and layouts can be encoded into embeddings if needed.



There exist several pre-trained models capable of producing pictures of sufficient quality for use in advertising. Such models include Dall-E (Ramesh et al., 2021, 2022) and the Stable Diffusion family of models (Rombach et al., 2021). Unfortunately, these models appear difficult to use directly in our application. Some models are proprietary and do not allow the user to directly navigate their embedding space, and other models have a very high dimensional embedding space (Stable Diffusion: 1024-dim., SDXL: 2048-dim.), which renders embedding space exploration difficult due to the curse of dimensionality.

The challenge at hand is thus to design and train a generative model where the embedding space is of low enough dimensionality to facilitate exploration, but where the quality of the output is sufficient for use in advertising. To achieve this, we build the generative model in two stages (see Figure 3). For the first stage, we train a model capable of generating low quality landscapes. The high specialization of this model (only landscapes) allows us to maintain an embedding space of

lower dimensionality. The lower quality requirements for the generated output allows us to keep a simpler architecture, use a smaller training sample, and save on computational resources. In the second stage, we feed this low quality picture to SDXL (Podell et al., 2023), a large pre-trained model, which can improve the quality to a degree acceptable for use in advertising (see examples in Figure 4).

Figure 4. Examples of generated landscapes.

The first-stage generative model combines two small generative architectures to keep the dimensionality manageable: an adversarially trained variational auto-encoder, or VAE-GAN (Khan et al., 2018; Larsen et al., 2015; Plumerault et al., 2021), and SPADE (Park et al., 2019). The VAE-GAN generates a layout image, indicating the location of various elements such as trees, rocks, and water, from a 48-dimensional embedding vector. The SPADE architecture converts this layout to a low quality (256x256 pixels) image using another 256-dimensional embedding vector defining the style (e.g., sunset, savanna type, etc.). As a result, any generated landscape can be summarized by a 304-dimensional vector combining the layout and style embeddings. After training, we can draw a location in this combined 304-dimensional

embedding space to generate a landscape. Both architectures are also capable of respectively encoding the layout and style from existing images into their embedding representations. The separation of landscape and style is necessary to allow us to fix minor issues with some images, such as trees floating in the sky. These issues are likely due to the limited size of the dataset used for training and the use of a simple architecture. By using a separate auto-encoder for the layout, these minor issues can easily be manually corrected in the landscape layout (see Figure 5). The corrected layout can then be fed back to the auto-encoder to obtain the updated embedding space location. If we used instead a single architecture generating landscapes directly from the embedding space, we would not have the ability to fix the layout.

Our first stage is trained on a dataset of 135,719 landscape images, where 45,719 were images tagged as landscapes on Flickr (via the official API) and 90,000 images come from the Landscapes High-Quality (LHQ) dataset (Skorokhodov et al., 2021). The layout is extracted from these images using the DeepLab V2 model (Chen et al., 2018). It is important to note that our training data has limitations. Specifically, our training data does not cover every possible type of natural scenery that exists and the elements of landscapes that our model can include in the layout images are also limited by the architectures of the used models. For instance, if an element is defined as "sky," the layout does not specify whether it refers to a night sky, a daytime sky, or a sunset sky. In this case, the variety of skies generated by the model is captured by the style vector and depends on the diversity of "sky" instances present in the training data. In our training data from Flickr and LHQ, shots of night skies were relatively scarce compared to daytime or evening skies. This scarcity of variation directly affects the model's output diversity.

Figure 5. Example of a corrected layout and the corresponding low quality landscapes generated in stage 1. The “tree” spot in the sky, represented in dark color on the layout, as been removed

	Original	Corrected
Layout		
Picture		

Once a low quality landscape picture is generated, an SDXL 1.0 model is used in the second stage to enhance the low quality images to high-quality 1024-by-1024 pixels images. SDXL 1.0 is a pre-trained state-of-the-art image generation model freely available on the Hugging Face platform. While there are several other models that could be used for this step, SDXL was chosen as it is free and provides us with landscape pictures of sufficient quality. We used the standard image to image pipeline with default settings.

Acceptability Prediction Model

Before generating landscape images to test them in the field, we need to be able to predict which images are acceptable for the brand. To achieve that, we train a neural network that maps the embedding representation of a landscape (both layout and style) to an acceptability rating between 0 and 1. The network is a series of fully connected layers using leaky ReLu activation functions. The network possesses 6 layers of sizes [304 128 64 32 8 1], with the output layer being fed to a sigmoid activation function of range 0-1. This architecture returned better results in simulation than smaller architectures, while requiring less training than larger architectures.

To train the model, we adopt a Bayesian active learning approach. This allows us to reduce the size of the training set, which is produced by having a company representative look at batches of landscapes and rate them as acceptable (1) or not (0). Our goal is to identify which batches of landscapes would be informative for the training. For instance, showing 10 nearly identical landscapes would not be informative, as only one of them would allow us to predict the likely acceptability rating of the nine others. Similarly, showing landscapes with low uncertainty around their acceptability rating (e.g., because they are similar to already rated landscapes), would not be very informative. We thus need an objective function measuring how informative a batch would be, and define a batch as being informative if the prediction probability that an image is rated high changes a lot after observing the actual ratings of the batch. In the Bayesian framework, this corresponds to the posterior predictive being highly different from the prior predictive. A traditional measure of this difference is the Kullback–Leibler divergence between these distributions, which also corresponds to the Shannon Information obtained with a given batch (Chaloner & Verdinelli, 1995). Formally, we denote the set of acceptable landscapes X^A,

and our prior beliefs about this set are represented by the distribution $p(X^A)$. We want to find the batch of landscapes X^B with acceptability ratings A^B that maximizes the divergence:

$$\int log\left(\frac{p(X^A|A^B,X^B)}{p(X^A)}\right)p(X^A|A^B,X^B)dX^A.$$

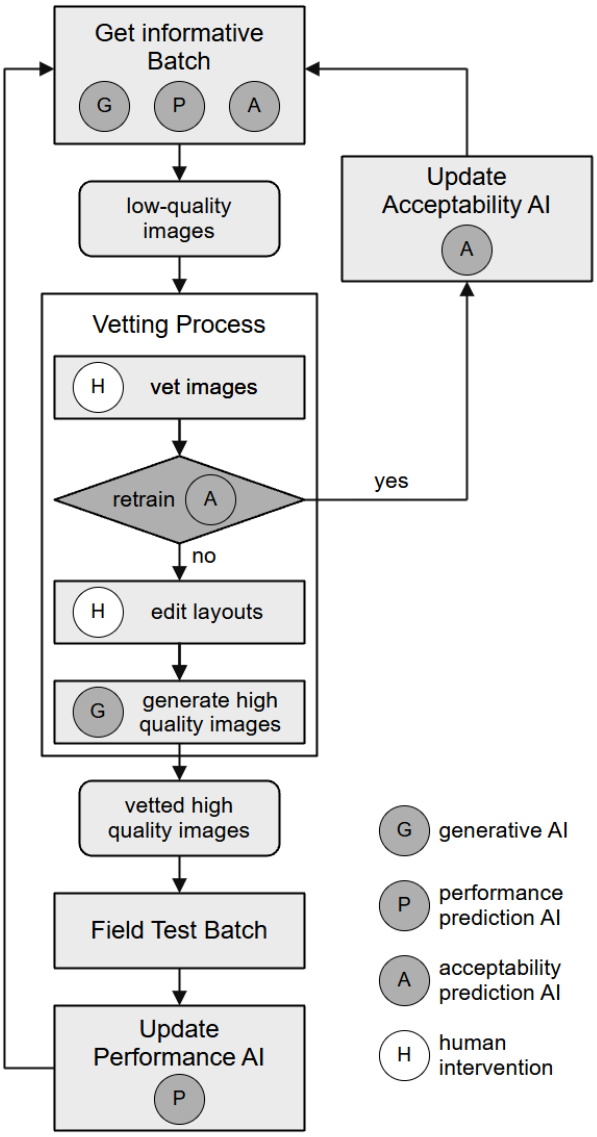
However, since we do not know the ratings before the batch has been manually labeled, it is usual to instead maximize the expected divergence where the beliefs about the ratings $p(A^B|X^B)$ are provided by our current model:

$$\int\left[\int log\left(\frac{p(X^A|A^B,X^B)}{p(X^A)}\right)p(X^A|A^B,X^B)dX^A\right]p(A^B|X^B)dA^B.$$

This process of finding an informative batch, getting the associated ratings, and using them to train the model is repeated several times, where the posterior distribution of ratings after a given iteration becomes the prior distribution for the next iterations. This approach of obtaining a sequence of distributions constitutes a typical sequential Bayesian filtering task. Computationally, the integration is done by Monte Carlo simulation, where a Hamiltonian Sequential Monte Carlo algorithm (Burda & Daviet, 2023) is used to simulate the posterior distribution. This algorithm is particularly suited to neural networks as it can leverage the automatic differentiation system integrated with most neural network libraries.

Performance Prediction Model

Figure 6. The process used to train our performance prediction AI.



Once equipped with a model capable of predicting whether an image is acceptable with sufficient accuracy, our next goal is to train a model predicting which images will achieve a high CTR. We follow the same approach as for acceptability prediction and use a neural network mapping the embedding representation of a landscape (both layout and style) to a predicted CTR ranging

between 0 and 10%. The same architecture is used, with 6 layers of sizes [304 128 64 32 8 1], but where the the output layer contains a sigmoid activation function of range 0-0.1.

To obtain the training data for this neural network, the images need to be integrated in an ad creative and tested in the field, which costs $10 to $20 for each creative tested. These images must also be vetted by a company representative to ensure that they satisfy the brand standards. In the case a batch does not satisfy these standards, the acceptability AI is updated with the ratings of that rejected batch. The training process is represented in Figure 6.

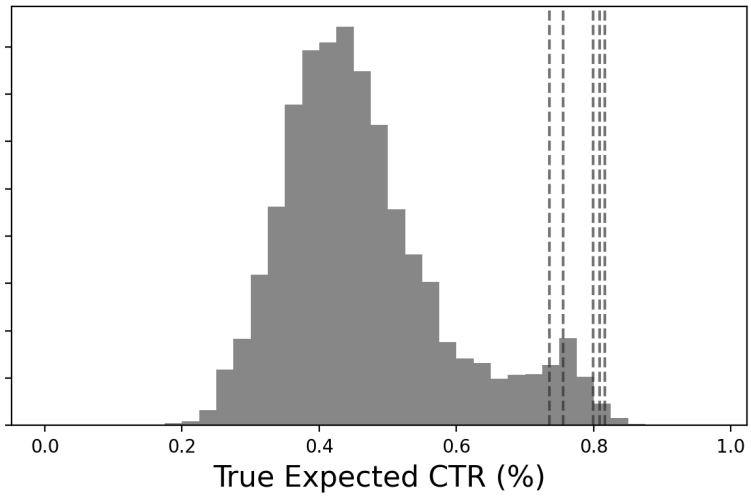
Our objective remains to find a batch X^B maximizing the expected Kullback–Leibler divergence between the prior and posterior beliefs on which landscapes X* achieve the top CTR. Since a result for a given creative can only be observed if its batch satisfies the brand standards, the probability of a batch passing the vetting process must be taken into account, and the objective function becomes:

$$\int \left[\int log \left(\frac{p(X^*|CTR^B, X^B)}{p(X^*)} \right) p(X^*|CTR^B, X^B) dX^* \right] p(A^B|X^B) p(CTR^B|X^B) dA^B dCTR^B.$$

We use the same Hamiltonian sequential Monte Carlo algorithm to simulate our distribution and perform Monte Carlo integration.

Performance in Simulation

Figure 7. Distribution of true expected CTRs and performance of the landscapes selected as best performer by our model in 5 separate simulations (dashed lines).

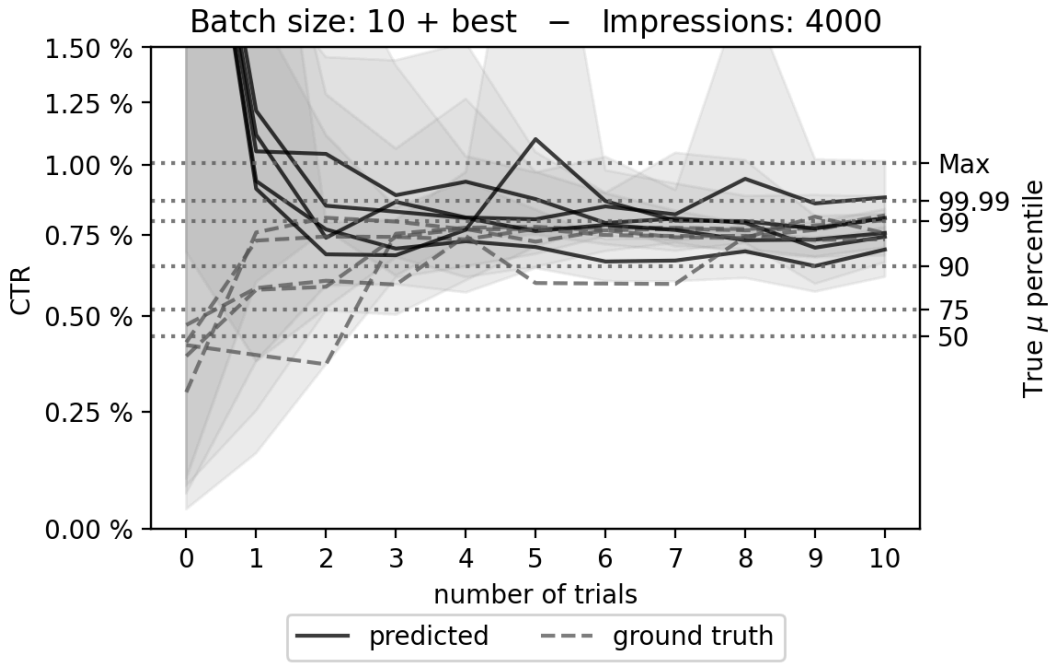


To test the optimization algorithm, we run a simulation where the expected performance of each landscape (ground truth) is known and test if our approach can identify high performing images rapidly. We use this simulation to test various network architectures (number and size of layers) and batch sizes. To compute the true expected CTR, we first randomly select a target landscape image in the training set with embedding representation X^* of dimension $K=304$ and confirm that it is acceptable with respect to brand standards. We then define the true expected CTR μ for any landscape X to decrease exponentially with the squared Euclidean distance to X^*:

$$\mu(X) = exp\left\{-4.6 - \frac{1}{K}\sum_{k=1}^{K}(x_k - x_k^*)^2\right\},$$

which has a maximum of 0.01 in X*. This defines the "ground truth" and results in a distribution where 99% of the landscapes in our training dataset have an expected CTR below 0.08 (see Figure 7).

Figure 8. Predicted performance (solid black lines) and true expected CTR (dashed lines) for the best predicted landscape after each trial. Each line is a different simulation.



We run 5 separate simulations for 10 trials with 4,000 impressions per landscape. The observed clicks are drawn from a binomial distribution, and we monitor the true and predicted CTRs for what the model under training predicts to be the best landscape. We test settings with batches ranging from 5 to 20 landscapes per trial. While we generally have good results with batches of 5, the variability in CTR across simulations leads to the model predictions converging slowly for some simulations. With batches of 10, all the simulations converge by the 5th trial. Increasing the batch sizes further does not lead to significant improvement while increasing the theoretical costs. We provide the output of a set of simulations with batch size 10 in Figure 8. We see that for most simulations, the model's predicted best landscapes fall around the 90th percentile or above starting with trial 3.

4. Field Experiment

Experimental Setting

Figure 9. Simplified representation of the template provided by the company applied to one of the generated landscapes.

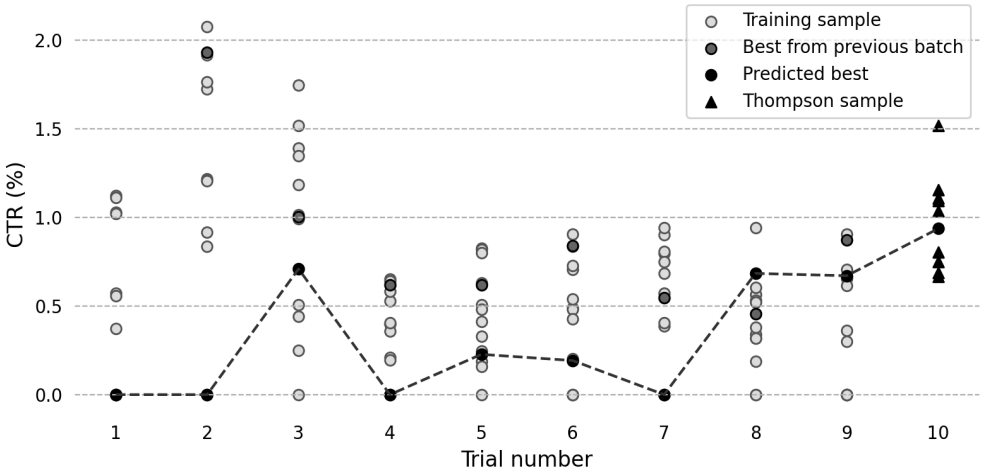
We now proceed to test our approach in a field experiment. First, the AcceptAI is trained with 60 batches containing 12 visuals, as well as the image predicted to have the highest acceptance probability (to verify prediction accuracy). The manual labeling took around 1.5 hours in total. To train the PerfAI, we need to test visuals in advertising campaigns in the field and measure the CTR. The partner company provided a template containing an elaborate shape with the main message in the center and a logo in the bottom right corner. We did not obtain the authorization to reproduce this design here and show instead a simplified representation in Figure 9. We apply this template on the vetted generated landscapes for each batch and use them in a week-long advertising campaign with a budget of $200 on the Instagram platform. Given our budget, we use optimized batches of 12 visuals, which include the visual predicted to perform best. Over the 9 batches tested for our experiment, a total of 108 visuals were produced for training the PerfAI.

The platform optimizes the budget allocation between creatives within each batch to attempt to maximize overall CTR, resulting in the best performing creatives having a lot more impressions than underperforming ones. In extreme cases, a creative can be completely dropped from the campaign after 100 impressions if it did not obtain a single click, while the budget is diverted to better performers. This platform behavior of providing more impressions to some ads does not represent a major issue as getting more accurate CTR estimation for top performers is beneficial for our approach.

A major issue however, resulting from the batches being tested sequentially, is the potential appearance of seasonal effects. To take these into account, we add a seasonal parameter linearly in the last neuron of our performance prediction model, before the final activation function. This parameter is estimated separately for each batch. To ensure identification, we add to each tested batch the best performer from the previous batch as a 13th visual. Note that since this visual is taken from the previous batch, it does not represent a new visual in our training dataset for PerfAI. By comparing the performance of an identical creative across batches, we can obtain an estimate of this seasonal effect and avoid affecting the performance prediction results. During a pilot study, we tested identical batches twice and found that while the CTR was highly variable, the rank of the creatives was mostly conserved. This supports the hypothesis of a seasonal effect affecting most creatives similarly, and that the variations in performance are not just noise.

Results

Figure 11. Observed performance for each batch, with the model-predicted best landscape shown in black. The last sample is selected by Thompson sampling (black triangles).

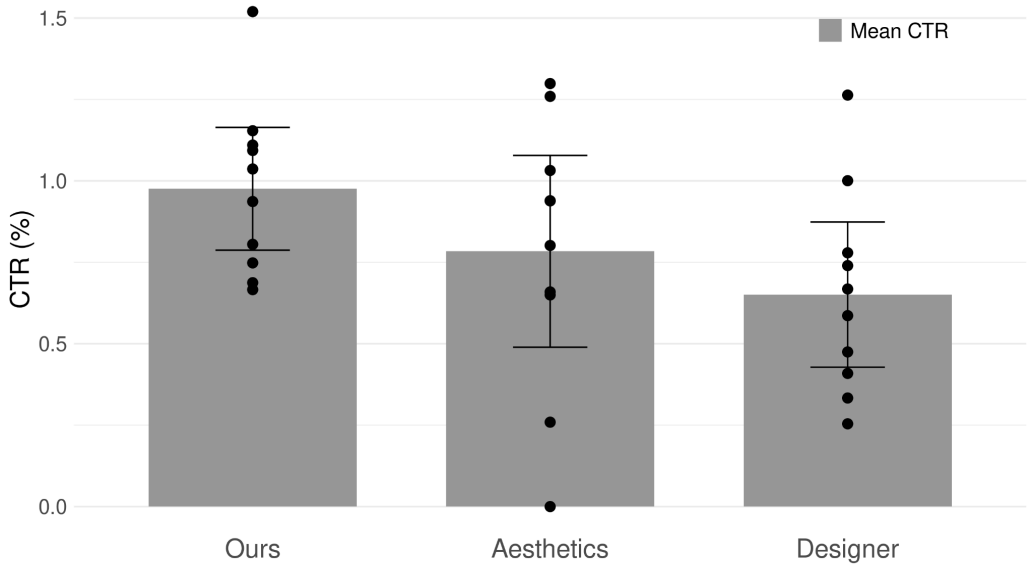


We report the CTR obtained for each batch in Figure 11. We can see heavy seasonal effects, and the data indicates that the same visual achieving a 2.1% CTR during a week (best performer at trial 2) can fall to 1.0% the next week when re-tested. Note that a major national holiday resulting in a long weekend falls during the testing of the second batch, explaining the seasonal effect. Looking at the creative predicted to perform best within each trial, we can see that for trial 1 to 7, the performance was mostly low, suggesting bad predictive abilities. The model only appeared to have learned to identify what a good visual could be starting trial 8.

After the last training batch, we evaluate the performance of our approach by combining nine visuals selected by Thompson sampling (Schwartz et al., 2017; Thompson, 1933) and the landscape predicted to perform best as a 10th visual. In the Thompson sampling approach, we sample the landscapes in the embedding space with a weight proportional to the probability that the given landscape has the highest predicted CTR according to our Bayesian model. This allows us to test the ability of our model to produce a batch of diverse landscapes that are likely to

perform well. We compare the performance of our Thompson sample to two other samples: 1) a sample of generated images predicted to have high aesthetic ratings by a pre-trained AI, and 2) a sample of images selected by a professional designer. These batches are tested concurrently to our Thomson sample and the results are reported in Figure 12. The images maximizing aesthetic ratings are produced by our generative AI using gradient descent to maximize the scores provided by TANet (He et al., 2022), a pre-trained AI showing a high performance in predicting human ratings of image aesthetics. The samples produced by the designer are just provided as a reference and we expect the performance to vary potentially broadly across designers. We have no claim that the creatives are the best possible human-produced designs.

Figure 12. Observed CTR performance of the sample selected by our model (left), the sample maximizing aesthetic ratings (middle), and designer-selected images (right). The error bars represent the 95% confidence intervals for the mean.



From a quantitative perspective, we note that the top achieved CTR in each batch are all between 1.25% and 1.50%. 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]. Looking at the CTR distribution, the images selected by our model have a higher average performance and less variation around that average, indicating that our model can reliably select good performers. From a qualitative perspective, the type of images selected by the designer were however more diverse. In particular, they contained a starry night, a field with flowers planted following a specific pattern, a pond under a canopy with rays of light filtering through the branches, and a recognizable geographical landmark (mountain). These pictures could not have been generated by our model, either because the landscape was not sufficiently represented in the training set (starry night, water under canopy), or because the patterns are too specific (flower patterns, landmarks). The other selected pictures could have reasonably been generated by our model and were similar to many pictures in our training set. We unfortunately cannot reproduce these images here for copyright reasons.

[2] https://www.xyzlab.com/meta-ads-benchmarks/japan reports CTR ranging from 1.3% to 1.9% depending on the industry at the time of writing (Oct 2024).

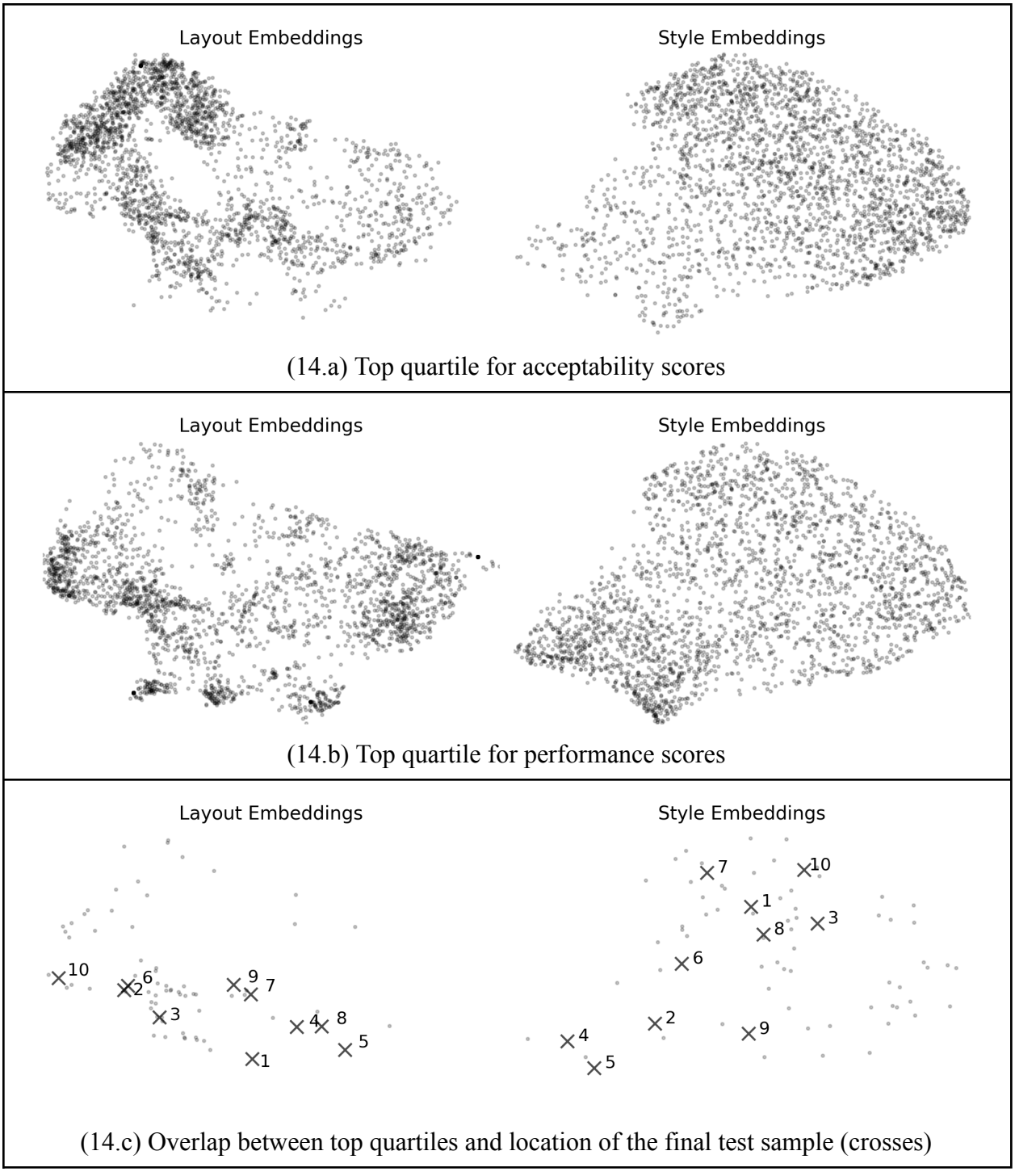
Visual Representation

Figure 13. UMAP representation of embedding vectors for the training data (dots) and tested visuals (squares).

To obtain a visual representation of the landscape space, we apply Parametric UMAP (Sainburg et al., 2021) to the training set's embeddings. This provides a 2-dimensional projection of the embedding space, where similar points in the embedding space remain close to each other in the projection space. The parametric version of UMAP also trains a neural network to learn the mapping from the original space to the projected one, allowing us to subsequently project the generated landscapes for each trial. Note that in our 304-dimensional embedding space, 48 dimensions represent the layout and 256 dimensions the style. A single UMAP projection would thus mostly capture similarities in style, with little weight put on the layout. When trying this approach, we indeed observed that points that were close to each other had similar textures, weather, and lighting, but drastically different geographies, such as a beach and a rocky mountain. As a result, we instead decided to project the layout and style embeddings to separate UMAP spaces (see Figure 13). We can see by looking at the projection of the batches that there

is a convergence towards specific areas of the layout space through the trials, whereas the style space remains well covered.

Figure 14. Predicted top quartiles for acceptability scores (top), performance scores (middle), and members of both top quartiles (bottom). The crosses indicate the location of the final sample.



(14.a) Top quartile for acceptability scores

(14.b) Top quartile for performance scores

(14.c) Overlap between top quartiles and location of the final test sample (crosses)

After training, we can predict which landscapes satisfy brand requirements using the AcceptAI and which should perform well using the PerfAI. We represent the location of the top quartiles for both acceptability and performance in Figure 14. We can see that the locations with a high density for good acceptability and good performance have little overlap, especially in the layout space. Specifically, only 5.3% of the landscapes in our training sample are in the top quartiles for both acceptability and performance (Figure 14.c). When retaining only top deciles, the overlap goes down to 0.6%. In other words, the kind of landscapes predicted to have a good performance rarely satisfy brand requirements. Given the difficulty to find good performers within the set of acceptable landscapes, using systematic methods like the one we propose appears to be very beneficial.

Figure 15. Set of creatives selected by Thompson sampling for final test.

We also project the embeddings of our final test sample containing the Thompson sampled visuals in Figure 14.c (identified by crosses) and provide the corresponding images in Figure 15 . We note that the creatives are dispersed in the embedding space and that various types of

landscapes are represented in the images, satisfying our need for diversity: sea shores, green hills, plains, mountains, and a waterfall. The type of climate represented is however similar across pictures, which is likely due to the acceptability AI restricting possible landscapes to fall within the styles covered in the company-recommended activities. By performing a reverse image search on Google[3], we were indeed able to find very similar landscapes in the geographic areas serviced by the company.

Further analysis reveals that several of the high performing landscapes are in areas with very few to no samples in the training data. For instance, one of these landscapes (image 8) has a sky to land ratio of 2:5, a mountain in the far-right background, a turbulent river flowing around numerous large rocks and boulders, dark evergreen trees, and a cloudy sky. An image with these specific characteristics was not present in our training set and might be difficult to locate in an image bank, either because it does not exist or because it has to be found among thousands of other images. This highlights the advantages of using a generative AI, allowing us to use visual compositions that would otherwise be difficult to obtain.

5. Discussion

In this paper, we have investigated how modern generative AI technologies for image synthesis could be leveraged in digital advertising. In particular we focused on the problem of identifying high performing visuals that are acceptable with respect to brand requirements while keeping usage of resources moderate. To achieve that, we proposed a novel approach combining a generative model with two other models respectively predicting acceptability with respect to brand requirements and CTR performance. These models were trained using a Bayesian active

[3] At the time of writing, Google Lens needs to be disabled with an extension to be able to perform a reverse image search supplemented by keywords such as geographical locations.

learning approach, allowing for the rapid identification of high-performing and brand-acceptable visuals.

The proposed approach was validated through both simulation and a field experiment in collaboration with a company promoting outdoor family activities. The results showed that our AI-generated visuals could rapidly achieve a performance comparable to a benchmark produced by a human designer. A further analysis of the high performing visuals revealed that many of them had no close counterpart in the large sample of real landscapes used to train the generative AI. This highlights the ability of the generative AI to compose new high-performing landscapes that might be difficult to obtain by a human designer. We noted however, that the human-produced visuals contained images that could have not been generated by our AI, either because they were very specific (e.g., recognizable landmarks) or because they fell out of our training data domain (e.g., starry night). We want to stress that our framework is not designed to surpass or imitate human abilities, as they are variable across individuals and possess different creative characteristics. We instead focus on maximizing the effectiveness of a generative AI system within its own creative capabilities.

While our approach shows promise, there are limitations. The quality and diversity of the generated visuals are inherently dependent on the training data and the capabilities of the underlying AI models. This limitation was observed in our field experiment, where the lack of certain types of images in the training data restricted the variety of visuals that could be generated. Furthermore, the effectiveness of our system is contingent on the accurate prediction of CTR and acceptability, which may vary significantly across different industries and target audiences. Finally, this approach provides little insight on why a visual might be a high performer. While some methods inspired by research from the field of AI explainability could be

developed as an extension to the current framework, this issue is left for future research to investigate.

For small companies with limited resources, our proposed approach offers a significant advantage. Traditional methods of creating ad visuals potentially require investment in human capital and resources, which may be prohibitive for smaller firms. This is especially the case if many visuals need to be produced for sequential A/B testing, as there are substantial costs in designing, testing, and qualitatively analyzing the results. The ability to generate diverse visuals rapidly and at low cost using a generative AI can democratize access to high-quality creative content with minimal resources. Moreover, the efficiency of the AI-driven approach allows the company to rapidly adapt to market trends and consumer preferences. For larger companies, with more abundant resources, the proposed approach can be used as a first proof-of-concept stage to explore specific creative approaches. Once the results of various creatives are obtained and high performing visuals are identified, the company can then provide this information to a professional creative team who will take over.

As AI capabilities continue to grow at a rapid pace, we expect to see increasing research on generative models and their applications. This paper is one of the first attempts to explore the potential of generative technologies in digital advertising. We hope that this will inspire researchers and practitioners alike to attempt integrating generative AIs to their work, in particular in the visual domain.

Bibliography

Azimi, J., Zhang, R., Zhou, Y., Navalpakkam, V., Mao, J., & Fern, X. (2012, October 29). Visual appearance of display ads and its effect on click through rate. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. CIKM'12: 21st ACM International Conference on Information and Knowledge Management, Maui Hawaii USA. https://doi.org/10.1145/2396761.2396826

Brock, A., Donahue, J., & Simonyan, K. (2018). Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1809.11096

Burda, M., & Daviet, R. (2023). Hamiltonian sequential Monte Carlo with application to consumer choice behavior. *Econometric Reviews*, *42*(1), 54–77.

Burnap, A., Hauser, J. R., & Timoshenko, A. (2023). Product Aesthetic Design: A Machine Learning Augmentation. *Marketing Science*, *42*(6), 1029–1056.

Chaloner, K., & Verdinelli, I. (1995). Bayesian Experimental Design: A Review. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *10*(3), 273–304.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). 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.

de Gregorio, F., & Windels, K. (2021). Are advertising agency creatives more creative than anyone else? An exploratory test of competing predictions. *Journal of Advertising*, *50*(2), 207–216.

Dew, R. (2024). Adaptive preference measurement with unstructured data. *Management Science*. https://doi.org/10.1287/mnsc.2023.03775

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

Goldfarb, A., & Tucker, C. (2011). Online display advertising: Targeting and obtrusiveness. *Marketing*

Science, *30*(3), 389–404.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144.

Gurzki, H., Schlatter, N., & Woisetschläger, D. M. (2019). Crafting extraordinary stories: Decoding luxury brand communications. *Journal of Advertising*, *48*(4), 401–414.

He, S., Zhang, Y., Xie, R., Jiang, D., & Ming, A. (2022, July). Rethinking image aesthetics assessment: Models, datasets and benchmarks. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. Thirty-First International Joint Conference on Artificial Intelligence {IJCAI-22}, Vienna, Austria. https://doi.org/10.24963/ijcai.2022/132

Johnson, G. A., Lewis, R. A., & Reiley, D. H. (2017). When less is more: Data and power in advertising experiments. *Marketing Science*, *36*(1), 43–53.

Khan, S. H., Hayat, M., & Barnes, N. (2018). Adversarial Training of Variational Auto-encoders for High Fidelity Image Generation. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1804.10323

Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. In *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1312.6114

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O. (2015). Autoencoding beyond pixels using a learned similarity metric. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1512.09300

Liu-Thompkins, Y. (2019). A decade of online advertising research: What we learned and what we need to know. *Journal of Advertising*, *48*(1), 1–13.

Maier, E., & Dost, F. (2018). The positive effect of contextual image backgrounds on fluency and liking. *Journal of Retailing and Consumer Services*, *40*, 109–116.

Manchanda, P., Dubé, J.-P., Goh, K. Y., & Chintagunta, P. K. (2006). The effect of banner advertising on Internet purchasing. *JMR, Journal of Marketing Research*, *43*(1), 98–108.

Park, T., Liu, M.-Y., Wang, T.-C., & Zhu, J.-Y. (2019). Semantic Image Synthesis with Spatially-Adaptive Normalization. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1903.07291

Plumerault, A., Le Borgne, H., & Hudelot, C. (2021). AVAE: Adversarial Variational Auto Encoder. *2020*

25th International Conference on Pattern Recognition (ICPR), 8687–8694.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., & Rombach, R. (2023). SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/2307.01952

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. *ArXiv*, *abs/2204.06125*. https://doi.org/10.48550/arXiv.2204.06125

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-Shot Text-to-Image Generation. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 8821–8831). PMLR.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2021). High-Resolution Image Synthesis with Latent Diffusion Models. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/2112.10752

Sainburg, T., McInnes, L., & Gentner, T. Q. (2021). Parametric UMAP embeddings for representation and semisupervised learning. *Neural Computation*, *33*(11), 2881–2907.

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

Schweidel, D. A., Bradlow, E. T., & Williams, P. (2006). A feature-based approach to assessing advertisement similarity. *JMR, Journal of Marketing Research*, *43*(2), 237–243.

Sisodia, A., Burnap, A., & Kumar, V. (2023). Automatic discovery and generation of visual design characteristics: Application to visual conjoint. *Available at SSRN 4151019*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4151019

Skorokhodov, I., Sotnikov, G., & Elhoseiny, M. (2021). Aligning latent and image spaces to connect the unconnectable. *IEEE International Conference on Computer Vision*, 14124–14133.

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.

Yoo, J., & Kim, M. (2014). The effects of online product presentation on consumer responses: A mental imagery perspective. *Journal of Business Research*, *67*(11), 2464–2472.