Ad Image Generation by the Latent Diffusion Model

Yohei Nishimura
ynishimura@wisc.edu

Ping-Han Chuang
pchuang4@wisc.edu

Gabriel Gozum
ggozum@wisc.edu

Aanandita Dhawan
dhawan6@wisc.edu

1. Introduction

Modern text-to-image models produce results that are indistinguishable from their actual counterparts. Our work focuses on exploring recently-introduced Latent Diffusion Models [12]. Using an experimental advertisement dataset [6], we aim to train a model capable of producing convincing advertisements given simple human-generated prompts.

Advertising images convey complex meanings and concepts to encourage people to recognize brands or take actions [15], making it a challenging subject of analysis given advertisements contain context that goes beyond the visual accuracy of the image on its own. The creation of advertisements stems from the interaction between the marketer, who is in charge of the concept based on consumer analysis, and the designer, who transforms the concept into an ad [9]. In terms of the optimization of ad images, however, collaboration is a bottleneck; it is essential for optimization to generate as many ad images as possible quickly and to collect a large amount of feedback from users. Cooperative human tasks are a bottleneck inhibiting the turn-around time for ad production.

Effective marketing advertisements mainly consist of 1) text, 2) visuals (images and videos), and 3) others (music, etc.). The main focus of our research is to apply deep generative image models to create novel advertisement images. Previous research surrounding advertisement image generation is dominated by research consisting of component-wise tasks, such as advertisement layout and optimal text placement within an image [10].

Entire advertisements have not been generated from scratch yet because it is difficult to generate believable high-resolution images and it is difficult generating images with concepts capable of evoking user actions. As example, if an advertisement image promoting a banana smoothie beverage is generated using a prompt such as 'a photo of banana smoothie,' a simple product image is generated. Though these images are photo-realistic, they are not enough to encourage people to take action. Figure 1 shows images generated by the Stable Diffusion [1].

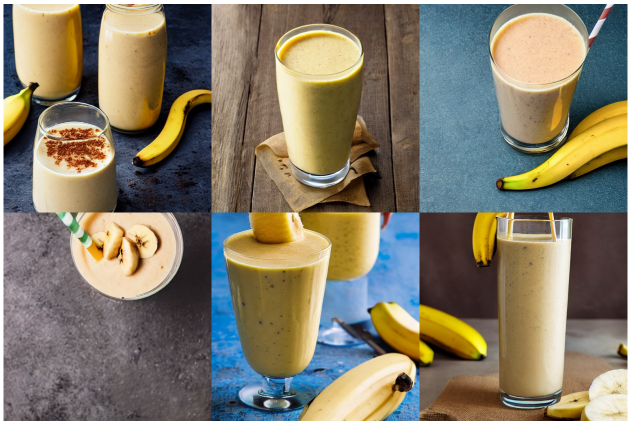


Figure 1. Images from stable diffusion model provided by huggingface.

In this study, we use recent image generation techniques combined with detailed and contextual prompts to generate a novel creative. As a result, this study contributes significantly to improving the accuracy and speed of ad optimization.

The trained embedding space by all topics and sentiments and all the images generated by our experiment is publicly released here.

2. Related Work

2.1. Marketing with Machine Learning

Research on marketing with advertisement images using machine learning is diverse. One central area is the prediction of advertising effectiveness; for example, [3] uses Convolutional Neural Network (CNN) algorithms to create predictive models of which online ads perform better. There are also many applications of computer vision research using deep learning; [17] builds the classification model of ad images using a CNN-based neural network.

In the field of ad generation, there are many component-wise types of research, such as the framework generation for banner images on websites [16], the extraction and generation of a logo by machine learning algorithm [2]. Additionally, there is research on the generation of advertising text (e.g., [8]), which applies Natural Language Processing

[1] https://huggingface.co/blog/stable_diffusion

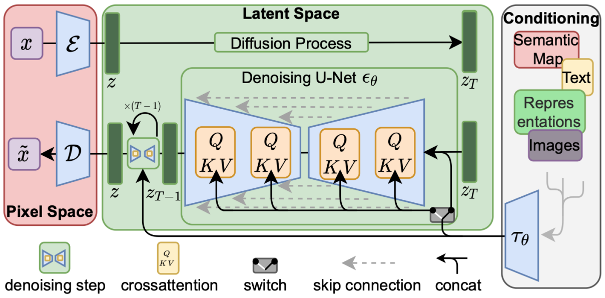


Figure 2. Latent diffusion model [12].

algorithms to the generation of ad sentences - this approach is actively used in industry.

2.2. Diffusion Model

Based on [7], diffusion models are an innovation in the world of deep learning. They are generative models in many domains, such as audio generation or image generation. A diffusion model is a parameterized Markov chain trained by variational inference to produce samples matching the data after a finite time. Transitions of this chain are learned to reverse a diffusion process, which is a Markov chain that gradually adds noise to the data in the opposite direction of sampling until signals are destroyed. When the diffusion consists of small amounts of Gaussian noise, it is sufficient to set the sampling chain transitions to conditional Gaussians, allowing for a particularly simple neural network parameterization.

3. Model

As background, we will introduce LDM that achieves state-of-the-art results for image generations tasks [12]. Additionally we will cover textual embeddings and textual inversion, both of which were used for our research.

3.1. Latent Diffusion Model

LDMs focus on improving the drawbacks of vanilla diffusion models. Prior diffusion models took full-images as input, requiring gradients to be calculated in the high-dimensional pixel space. Unreasonable amounts of compute power were required to fully train these models, on the order of thousands of GPU hours [11]. This large amount of compute placed diffusion models out of reach from the majority of researchers.

This is where LDMs come into play. As the name implies, LDMs first encode an inputted image into the latent space. This latent space can be thought of as lower dimensionality, removing high-frequency noise or imperceptible details from the original image. As shown in Figure 2 after encoding the input image a forward diffusion pass is ran, followed by a backwards. However, the backwards pass includes conditioned cross-attention at each stage of the typical denoising U-Net. Lastly, a denoising step is performed, followed by a decoder that converts the latent vector back into the image space, producing realistic images on output.

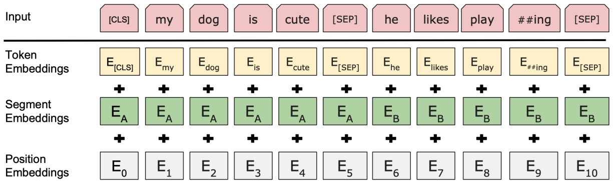


Figure 3. BERT tokenizer [1].

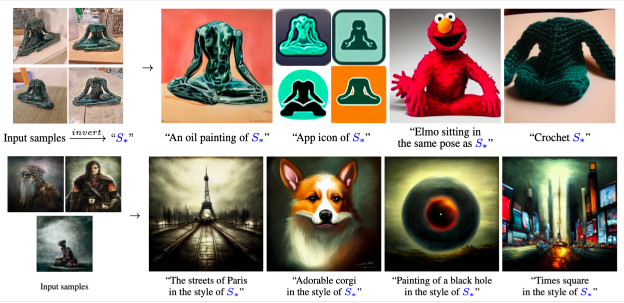

Input samples $\xrightarrow{invert}$ "S_*"
"An oil painting of S_*"
"App icon of S_*"
"Elmo sitting in the same pose as S_*"
"Crochet S_*"
Input samples
"The streets of Paris in the style of S_*"
"Adorable corgi in the style of S_*"
"Painting of a black hole in the style of S_*"
"Times square in the style of S_*"

Figure 4. Usage of textual inversion [4].

3.2. Text Embeddings

The BERT tokenizer is used to convert textual data into vectorized form [1]. As depicted in Figure 3, an input sentence is first split into individual words and subwords. These sub-slices are then converted into tokens - token embeddings can be thought of as a static key-value lookup. Lastly, a learnable segment embedding is added to the token, along with a positional embedding. Positional embeddings are typically sinusoidal functions, they give the model context as to the relative location of each token.

3.3. Textual Inversion

With the pre-trained LDM for text-to-image generation, the naive way to learn new concept is to fine-tine LDM, but it has some problems. First, it requires large amount of computing resources. Also, it hurts the model's capability to generalization. Thus, we adopt textual inversion [4] technique to let the model learn new concepts.

Textual inversion uses three to five images containing a new concept to learn a pseudo word representing the concept. Then, the pseudo word can be used to compose a sentence to prompt the model for generating image with the new concept. Concepts being learned can be concrete, like an specific object, or abstract, like the style of images. Figure 4 is an example.

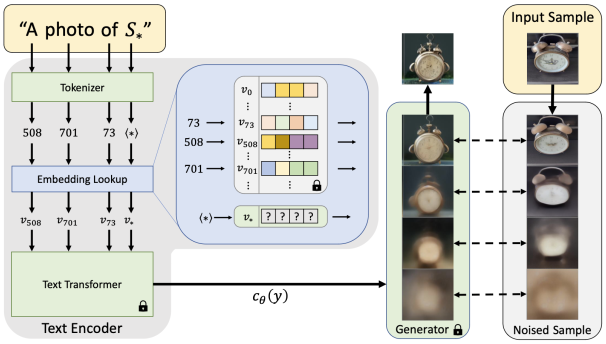


Figure 5. Textual inversion process [4].

As for how textual inversion learns a new pseudo word for a new concept, its goal is to change the embedding space of the text encoder so that pseudo word gets its word embedding v_*. Figure 5 shows the process of textual inversion. In the text encoder, it has a learned dictionary to map each token to a unique word embedding vector and that is the target textual inversion tries to extend. By giving three to five images about new concept and some context sentences as prompts, texture inversion goes through LDM's optimization process with the ϵ_θ and c_θ fixed. The loss function it use is the same as LDM, which is showed in equation 1, except that the modified term is word embedding v_*.

$$v_* := \arg\min_v \mathbb{E}_{z\sim\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t} \|\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))\|_2^2 \tag{1}$$

For hyperparameters used in textual inversion, we set the base learning rate to 0.005, the timesteps of LDM to 1000, and the scale factor of LDM to 0.18215.

4. Experiment

4.1. Overview

We fine-tune the embedding space of the Latent Diffusion Model with an ad-specific dataset using the Textual-Inversion method. The ad-specific dataset is divided into 37 sets along with topics and 30 sets along with sentiments using the annotation data. We fine-tune the embedding space by each topic and sentiment.

We evaluate generated results by Inception Score (IS) [13] and Fréchet inception distance (FID) [5], and manual inspection since only IS and FID can not assess the quality of the images as an advertisement.

4.2. Dataset

Figure 6. A sample image [6]

We use an experimental advertisement dataset [6]. The dataset can be found here. Table 1 reports a summary of the dataset consisting of 64,832 image ads. For example, in the ”sentiment” subset, a sample question is ”What is the sentiment of this advertisement?” for the sample image 6, and possible answers are ”creative,” ”eager,” ”inspired,” ”persuaded,” and ”youthful.” Our task is to find the optimal answer given the image and question.

Type	Count	Example
Topic	204,340	Electronics
Sentiment	102,340	Cheerful
Q+A	202,090	I should bike because it's healthy.
Symbol	64,131	Danger (+ bounding box)
Strategy	20,000	Contrast
Slogan	11,130	Save the planet... save you.

Table 1. Summary of the image dataset [6]

5. Implementation Details

First we pre-process the dataset, then we create topic and sentiment directories for each image, along with annotation data. Furthermore, we train the embedding space with each individual topic and sentiment. These embedding spaces are merged into a single and have placeholders assigned to each topic/sentiment.

Pre-processing Because the images in the dataset are different resolutions, we resized the images into $(W, H) = (512, 512)$ along with the input size for the latent diffusion model.

Division of images by topics and sentiments After pre-processing, we divide all images into each topic and sentiment group along with the annotation data. Note that, in the annotation, some images are assigned to multiple topics and sentiments. We classify an image into multiple groups if it is allocated to various topics/sentiments.

Training/merging Embedding Spaces We encode the concept of topics and sentiments into an intermediate representation of a pre-trained text-to-image model [4]. Along with the procedure in the previous section, we train the embedding spaces with each topic and sentiment, respectively. Through this process, we obtain 37 topic-trained embedding spaces and 30 sentiment-trained spaces. Finally, we merge them into one embedding space assigning different placeholders to differentiate each of the smaller embedding spaces.

Note that, we use the model from [12] using 1.4 billion parameter text-to-image model, which is trained on the LAION-400M database [14].

Sampling We test some of the prompts, such as 'a photo of an advertisement of TOPIC with SENTIMENT' and 'an online display ad of TOPIC in the style of SENTIMENT'. Based on our manual inspection, we use the prompt standardized with 'an advertisement of TOPIC in the style of SENTIMENT.' This prompt is based on Appendix D [4].

6. Results

The following section provides visual results generated by the trained embedding spaces and the pre-trained latent diffusion model.

6.1. Results with the combinations of each topic and sentiment

First of all, we show the results from the merged embedding space trained by all topics and sentiments. 11a and 8b show successful results while 8c show failure cases.

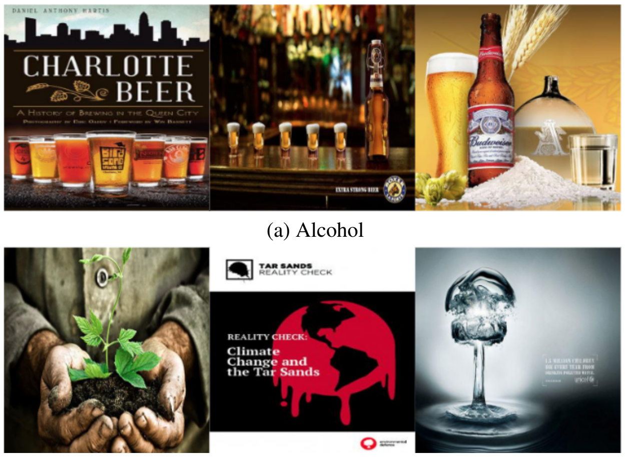
Based on the results, the model can generate ad-realistic results if the topic is 'concrete,' such as 'car' or 'alcohol.' In contrast, the model generates ambiguous images when the topic is 'conceptual,' examples being 'self-esteem' and 'environment.' The analysis is reasonable because the promotion of cars or alcohol encourages people to purchase the products. Also, the knowledge of how to promote is accumulated over many decades. As a result, the ad images for cars or alcohol mainly focus on the product picture, which is less complex of a task for the model to generate realistic images.

On the other hand, if marketers enlighten people to save the environment or inspire them to feel self-esteem, they tend to create stimulating, fascinating ads which indirectly deliver messages. Consistently, the images look imaginative, subjective, and abstract.

Figure 10 displays a sample image classified as 'alcohol' and 'environment' within the dataset.

Also, we create creatives for the car topic varying the scale parameter. Figure 9 shows the car-topic results with various scale parameters. We recognize that the results become refined and look realistic with the scale parameter increased while the diversity is lost; on the other hand, the small scale parameters generate a variety of advertisement images in spite of creating some unrecognizable results.

By manual inception, the results from car-and-fashionable combination with scale parameter 10 produces images of similar quality to existing catalogs or newspaper ad images - however, letters still appear abnormal.



(a) Alcohol

(b) Environment

Figure 7. Samples from dataset.

6.2. Results with specific images

Assuming practical usage of the technique, marketing experts may have their brands and specify the channel they use, such as posters or online display ads. In these cases, they can determine the ad images suitable for their products and channel.

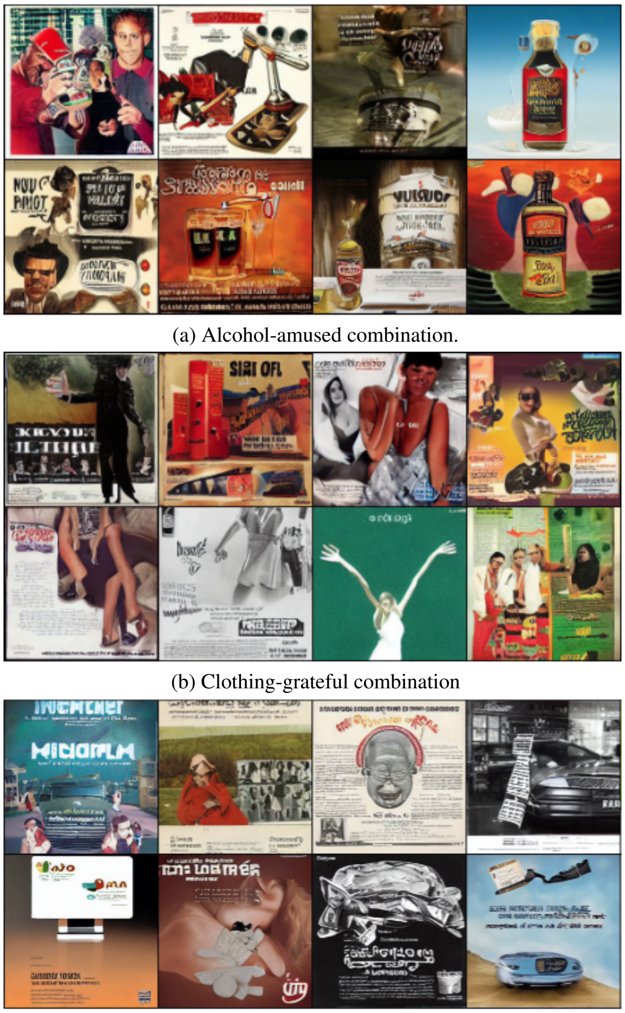
Based on this procedure, we experiment with the specific case for Starbucks promotion. Within the given dataset, we specify the images (see 10a) for Starbucks and train the embedding spaces with the three images.

We train the embedding space with three images and generate the ad-specific images by the prompt 'an advertisement in the style of TOPIC' (TOPIC is the placeholder for the embedding space).

Based on the specific condition, we can obtain the ad-realistic images by textual inversion technique. Figure 10a shows the generated results. We change the scale parameter to see change the diversity of the results. By manual inception, setting the scale equal to 10 resulted in the most realistic advertisements; however, small details in the logo differ from the actual.

Additionally, we perform ablations on intertwining Starbucks with other sentiments (active). To add the 'classic' Starbucks advertisement style, we select three Starbucks images (see 10b) and train the embedding spaces with the prompt 'an advertisement of TOPIC in the style of SENTIMENT.'

Figure 12 displays generated results with scale set to 10. Interestingly, the bottle's logo is replaced by the Starbucks logo (original being 7up). Ultimately, the more specific the base data, the more specific we can mimic a specific advertisement.



(a) Alcohol-amused combination.

(b) Clothing-grateful combination

(c) Healthcare-calm combination

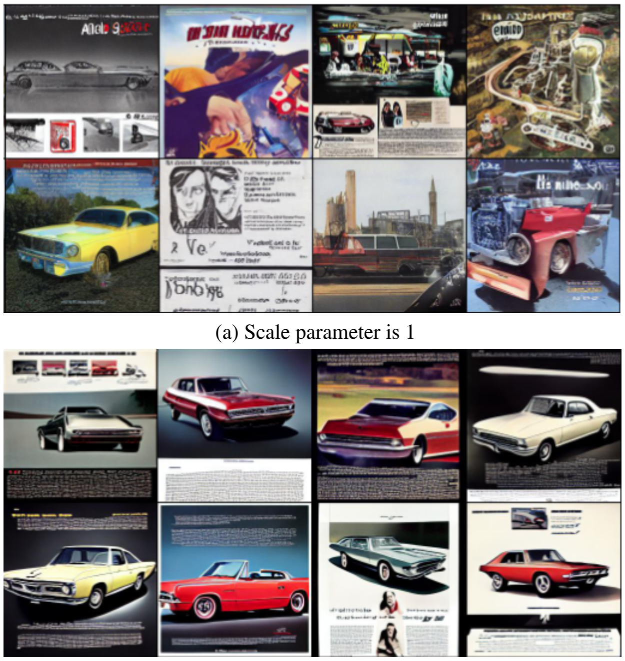
Figure 8. Various generated results with the scale parameter is 2, and the step size is 100.

6.3. Metric

Table 3 shows results for IS and FID metrics, results contain all topics and sentiments with step size set to 100 and scale parameter ablated on. The baseline is calculated using images generated from a prompt without placeholders; for instance, we use 'an advertisement of cars in the style of amazed.' We generate images with the same combinations as those with the placeholders.

The scores show that the generated images with placeholders are more diverse. Note that The scale parameter is set to 2 and step size to 100 for the baseline and all results. In terms of car results, FID is decreased with scale parameters smaller.

IS and FID of each category are shown in table 3; results place healthcare as the highest scoring, this agrees with our manual inspection of quality as well. Healthcare results are shown in Figure 8c.



(a) Scale parameter is 1

(b) Scale parameter is 10

Figure 9. Generated Images of the car-fashionable combination with different scale parameters. The step size is 100 in both experiments

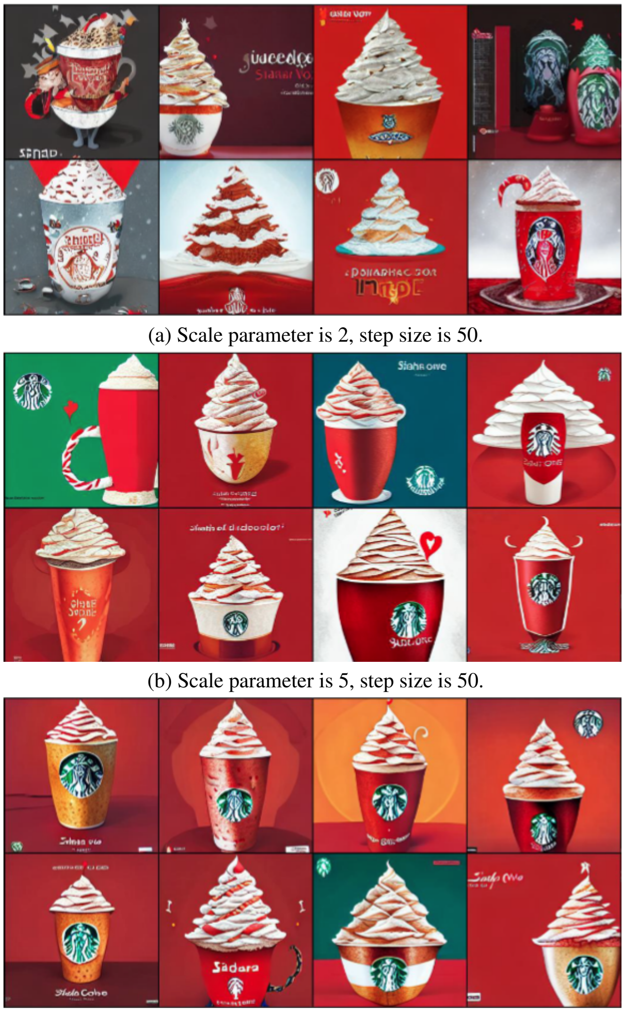
(a) Starbucks-spacific images to be used in training.

(b) Classic (in 'active' sentiment dataset) images to be used in training.

Figure 10. Images to be used in Starbucks-specific experiment.

7. Conclusion

In this paper, we introduce a holistic approach to generating images for advertisement. Using a concrete topic, we obtain realistic creatives, while we gain questionable generated images with conceptual (or general meaning) advertisement images. Moreover, we also generate images using



(a) Scale parameter is 2, step size is 50.

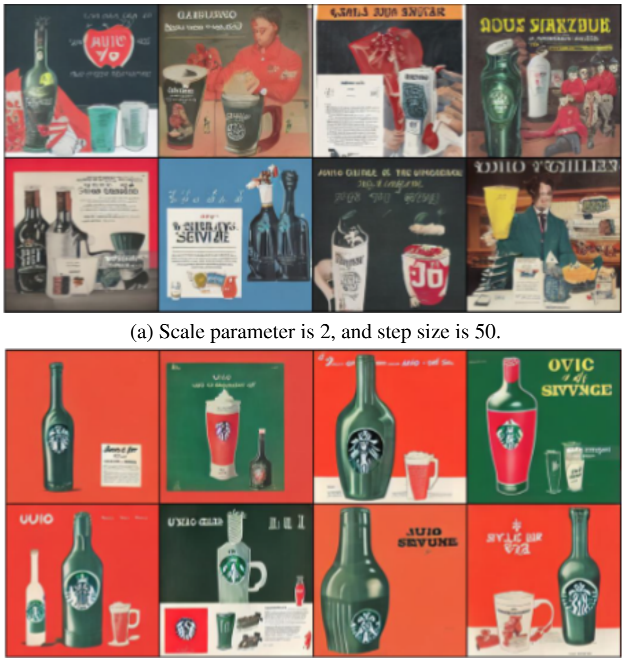
(b) Scale parameter is 5, step size is 50.

(c) Scale parameter is 10, step size is 50.

Figure 11. Results from Starbucks-specific embedding space.

category	IS	FID
baseline	**5.14**	36.54
All results	4.85	**27.34**
Car s=10	2.39	169.38
Car s=5	2.49	154.63
Car s=2	3.46	123.14
Car s=1	**4.72**	**106.42**

Table 2. ISs and FIDs of baseline, all results, and car results



(a) Scale parameter is 2, and step size is 50.

(b) Scale parameter is 10, and step size is 50.

Figure 12. Results from Starbucks-with-active embedding space.

category	IS	FID	category	IS	FID
healthcare	3.39	**90.95**	phone	3.06	99.73
financial	3.48	91.41	electronics	3.09	100.4
cleaning	4.29	92.18	gambling	3.2	100.59
shopping	3.58	92.53	alcohol	3.31	100.62
charities	3.61	92.59	chocolate	3.88	100.63
software	3.22	92.73	coffee	3.81	102.18
soda	3.56	92.94	home	3.11	104.17
sports	4.12	93.05	domestic	4.38	104.25
smoking	3.44	93.65	political	3.02	105.09
education	3.46	94.21	game	3.11	105.47
media	3.32	94.31	animal	3.65	107.32
human	3.51	97.17	seasoning	3.71	108.95
environment	4.13	97.95	baby	3.02	109.78
security	3.22	97.99	chips	3.59	111.44
clothing	4.03	98.1	travel	2.62	118.57
safety	4.35	98.24	cars	3.52	124.08
self	3.26	98.33	petfood	**4.72**	131.38
beauty	4.15	98.87	restaurant	3.27	133.92
home	4.21	99.49			

Table 3. ISs and FIDs of each category

an embedding space trained with specific company images and specific styles; results show generally realistic images with small inaccuracies in details, such as logos.

In practice, with using a specific and realistic dataset, the model can be used to generate new images for advertisement. Given the advertising dataset in our experiment is experimental, it includes a mixture of data from multiple channels (i.e. posters, newspaper ads, package design, and digital image ads). For better results in practice, the model should be trained with the data concentrating on a specific channel and product/service.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. 2

[2] Ryan Dew, Asim Ansari, and Olivier Toubia. Letting logos speak: Leveraging multiview representation learning for data-driven logo design. *SSRN Electronic Journal*, 2019. 1

[3] Michael Fire and Jonathan Schler. Exploring online ad images using a deep convolutional neural network approach, 2015. 1

[4] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. 2, 3, 4

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, abs/1706.08500, 2017. 3

[6] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1705–1715, 2017. 1, 3

[7] Pieter Abbeel Jonathan Ho, Ajay Jain. Denoising diffusion probabilistic models. 2

[8] Hidetaka Kamigaito, Peinan Zhang, Hiroya Takamura, and Manabu Okumura. An empirical study of generating texts for search engine advertising. In *NAACL*, 2021. 1

[9] Scott Koslow, Sheila L Sasser, and Edward A Riordan. What is creative to whom and why? perceptions in advertising agencies. *Journal of advertising Research*, 43(1):96–110, 2003. 1

[10] Hsin-Ying Lee, Weilong Yang, Lu Jiang, Madison Le, Irfan Essa, Haifeng Gong, and Ming-Hsuan Yang. Neural design network: Graphic layout generation with constraints. *CoRR*, abs/1912.09421, 2019. 1

[11] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 2

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752, 2021. 1, 2, 4

[13] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 3

[14] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021. 4

[15] Demetrios Vakratsas and Tim Ambler. How advertising works: What do we really know? *Journal of Marketing*, 63(1):26–43, 1999. 1

[16] Sreekanth Vempati, Korah T. Malayil, Sruthi V, and Sandeep R. Enabling hyper-personalisation: Automated ad creative generation and ranking for fashion e-commerce. *CoRR*, abs/1908.10139, 2019. 1

[17] An Tien Vo, Hai Son Tran, and Thai Hoang Le. Advertisement image classification using convolutional neural network. In *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*, pages 197–202, 2017. 1