Injection of Cultural-based Subjects into Stable Diffusion Image Generative Model

Amirah Alharbi[1] *amnharbi@uqu.edu.sa*

Reem Alluhibi[*,1] *s441007375@st.uqu.edu.sa*

Maryam Saif[1] *s441003907@st.uqu.edu.sa* Nada

Altalhi[1] *s441008451@st.uqu.edu.sa* Yara

Alharthi[1] *s441010323@st.uqu.edu.sa*

[1]Department of Computer Science and Artificial Intelligence, College of Computing,

Umm Alqura University, Makkah, KSA

Abstract

While text-to-image models have made remarkable progress in image synthesis, certain models, particularly generative diffusion models, have exhibited a noticeable bias to-wards generating images related to the culture of some developing countries. This paper introduces an empirical investigation aimed at mitigating the bias of image generative model. We achieve this by incorporating symbols representing Saudi culture into a stable diffusion model using the Dreambooth technique. CLIP score metric is used to assess the outcomes in this study. This paper also explores the impact of varying parameters for instance the quantity of training images and the learning rate. The findings reveal a substantial reduction in bias-related concerns and propose an innovative metric for evaluating cultural relevance.

Keywords:

Generative model; Diffusion model; Bias; Saudi culture; text-to-image

1. Introduction

The first appearance of generative models indirectly was in 1986 by introducing the concept of backpropagation [1], which enabled us to train neural networks, which is an essential part of current generation models. The author introduced the Back-propagating Error algorithm which is the basis of neural networks and machine learning, therefore it considers as a core component of training generative models. Generative models are a type of deep learning. The concept behind generative mod-els revolves around their ability to create new data by learning the underlying dis-tribution of the dataset they were trained on. They leverage deep neural networks to understand complex representations of the data. Generative models may also per-form tasks such as image processing, voice recognition and information retrieval as mentioned in [2]. Generative models have been proven to generate images [3], videos, texts, and music [4], [5], [6], [7]. They also have the ability to generate high-quality,

realistic outputs. Thus, generative models can be employed in several areas such as games, music production and design. However, there are still limitations and room for improvement [8].

The categories within the realm of image generation models encompass Image-to-Image translation models, which are oriented towards translating images from one domain to another, and layout-to-image generation models. The latter category focuses on generating images based on prescribed layouts or structural arrange-ments.ses on generating images based on prescribed layouts or structural arrange-ments. These models have found application in diverse domains, including object generation, scene synthesis, and video sequence generation.

Text-to-Image generation models can be classified into distinct categories, which encompass Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Conditional GANs (cGANs). VAEs are regarded as probabilistic mod-els designed to acquire a latent representation of input data, facilitating the gener-ation of novel samples through sampling from this latent space. In 2014 with the introduction of the GAN (Generative Adversarial Network) model, which is an an-tagonistic generation model that can generate high-quality images [9]. It was the first method in generative modeling that achieved impressive results. Since then, other generative models have been developed, such as VAE (Variational Autoencoder) models and Flow-Based models. GANs, on the other hand, comprise both a gener-ator network responsible for generating new samples and a discriminator network tasked with distinguishing between real and generated samples. The training process of GANs involves an adversarial interplay between these two networks. Conditional GANs (cGANs) represent an

extension of the GAN framework by incorporating additional input information, such as class labels or textual descriptions, into the generation process. This conditioning allows for more controlled and context-aware image generation.

Image generation models can be classified into distinct categories, which encom- pass Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Conditional GANs (cGANs) [10]. VAEs are regarded as probabilistic models de- signed to acquire a latent representation of input data, facilitating the generation of novel samples through sampling from this latent space. GANs, on the other hand, comprise both a generator network responsible for generating new samples and a dis- criminator network tasked with distinguishing between real and generated samples. The training process of GANs involves an adversarial interplay between these two networks. Conditional GANs (cGANs) represent an extension of the GAN frame-work by incorporating additional input information, such as class labels or textual descriptions, into the generation process. This conditioning allows for more con- trolled and context-aware image generation. Additional categories within the realm of image generation models encompass Image-to-Image translation models, which are oriented towards translating images from one domain to another, and layout-to-image generation models. The latter category focuses on generating images based on prescribed layouts or structural arrangements. These models have found appli-cation in diverse domains, including object generation, scene synthesis, and video sequence generation [10].

Diffusion models(DMs) or (Diffusion Probabilistic models) belong to the category of generative models. It uses Markov chain trained using variational inference to produce samples matching the data after finite time [11], which have shown remark- able performance in various domains, including the text-to-image art. These models have demonstrated competitive results that rival the state-of-the-art GANs (Gener- ative Adversarial Networks) commonly used in the field. Their effectiveness lies in their ability to generate high-quality images based on textual input, making them a valuable tool for artistic expression and creativity [12]. Further description of the architecture of diffusion models are provided in the next section. Stable Diffusion [13] v1.4, v2, Imagen [14] and DALL-E [15] are examples of Diffusion-based text-to- image (TTI) systems that are one of the most recent machine learning approaches in prompt image generation. Figure 1 we show the timeline of the presence of dif- ferent diffusion models. Generative AI models, such as text-to-image models, can suffer from biases due to their reliance on large

datasets that may contain biased or degenerated human behavior, which is the presence of unfair or unjustified pref- erences or prejudices in the output of generative text-to-image models. Training datasets obtained through web scraping can introduce biases, which encompass but are not limited to harmful, pornographic, mislabeled, and corrupted examples. The process of filtering explicit content from training data can inadvertently exacerbate bias-related issues, necessitating the implementation of supplementary measures for bias mitigation. Consequently, text-to-image systems such as Stable Diffusion re- liant on web-scraped training data, filtering procedures, and the incorporation of guiding models are susceptible to the propagation of bias. It is imperative that further research endeavors are undertaken to comprehensively comprehend and ef-fectively address these biases [16]. In our investigation, we noticed that the Stable Diffusion model suffers from bias against the Saudi culture and prompted us to dis- cover the effectiveness of dreambooth technique to inject the Saudi culture subjects to diffusion models.
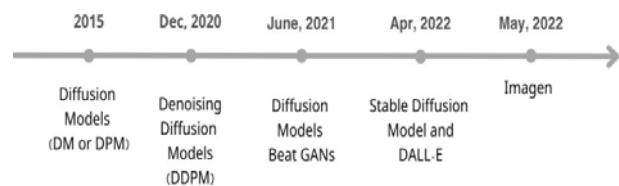

Fig. 1.   Evolution timeline of diffusion models

1.1 Motivation

This paper is originated from a local project to support the ministry of municipal and rural affairs to remove visual distortions in streets such as graffiti. We aimed to harness generative AI technique for generating images that can reflect Saudi local culture and then can be used as a replacement of the distortions on walls streets. The realization of the pervasive issue of bias within generative models and the consequential inadequacies in generating models reflective of Arab culture, has be- come increasingly evident and significant. Hence, the incorporation of Arab cultural into generative models increases their comprehensiveness, universality, and notably enhances the efficacy of the model within the specific cultural context of Saudi Arabia—this serves as the foundation of our initiative. This prompted us to look for different ways to address this bias and the dreambooth technique is a state of art option to discover its effectiveness in

injecting the Saudi culture subjects into diffusion models.

This paper is organised as follows. First, some related work of image generative mod-els are highlighted. Then research method pipeline is discussed elaborately along with the used algorithms for generating artistic images. The detailed experiments are demonstrated. After that, the results of the experiments are presented with different parameters included in our ablation study. Finally, a discussion of theresearch findings and a conclusion of this research are provided.

2. Related Work

Deep learning has propelled computer vision and image synthesis to unprecedented heights, with generative imagery playing a pivotal role in these advancements. This section dives into the multifaceted area of generative imagery, tracing its evolution from fundamental principles to state-of-the-art methodologies. Moreover, it delves into a critical ethical dimension of deep learning, shedding light on the pervasiveissue of bias and its implications for artificial intelligence (AI) systems. This work

[17] provides an extensive survey of image synthesis techniques tailored to the do- main of visual machine learning, systematically categorizing them based on their unique modeling and rendering approaches. By doing so, it offers readers a well- structured understanding of the capabilities of each technique. A central focus of this paper revolves around the growing importance of synthetic data within the complex training pipelines of deep learning applications. It emphasizes the rising potential of synthetic data, positioning it as a critical component that enhances the training process and increases its effectiveness.

This survey places particular emphasis on the intricate aspects of computer graphicsembedded within these methodologies. It highlights the role of computer graphics in shaping the evolving landscape of image generation for machine learning, revealing the significant intersection between the domains of computer graphics and machinelearning. This intersection signifies a promising path for transformative advance- ments in this dynamic field. As a significant contribution, this paper conducts a rigorous evaluation of each image synthesis method, employing a discerning lens to assess their quality and reported performance. These evaluations, firmly grounded in empirical evidence, offer valuable insights into the learning potential inherent in these image synthesis techniques.

Also, this paper [10] offers a comprehensive evaluation of the current landscape of im- age generation methods, systematically categorizing them based on the algorithms utilized, the data types they leverage, and their primary objectives. In a thorough and methodical manner, the authors provide in-depth explanations of each image generation technique, delineating their distinctive features and introducing their proposed approaches within each specific category.

This work explores the inherent limitations of select state-of-the-art methodolo- gies, digging deeper into some considerations such as the materials employed, inputdimensions, and training duration for each model. This analysis contributes to a comprehension of the practical implications of these methods. To assess the effec- tiveness of these solutions, the authors apply established metrics, including the In- ception Score (IS), Frechet Inception Distance (FID), and the Structural Similarity index measure (SSIM), offering a robust quantitative evaluation of their perfor-mance.

For a comprehensive illustration of their findings, the paper presents experimental results, featuring state-of-the-art methods evaluated across standard image genera- tion datasets, including DeepFashion, Market-1501, MNIST, CelebA, and COCO- stuff. This contextualizes the research within real-world scenarios, providing a tan- gible basis for their analysis. By comparing the performance of existing image gen-eration methods, the paper sheds light on the complicated interplay between the datasets employed and the requisite training time, furnishing valuable insights into the trade-offs and strengths inherent in this multifaceted domain Then, this paper

[18] introduces the innovative Retrieval-based methods RDM model, Retrieval-based methods in image generation models encompass the utilization of existing images or text data to generate new images. These techniques harness the capabilities of advanced models like CLIP [19] for retrieval and conditioning purposes. By incorpo- rating pre-existing images or text as conditioning inputs, retrieval-based methods enhance the image generation process, ultimately leading to improved relevance andaccuracy in the generated images.

Subject-driven generation, on the other hand, refers to the process of generating images based on specific concepts or subjects provided by the user. This approach empowers a more interactive and personalized image generation experience, where users can input precise images that encapsulate their desired concepts or themes. The key advantage is the ability to generate

images that closely align with the user's intended subjects or themes.

Various methods have been proposed to integrate subject-driven generation into image generation models. These methods frequently involve the use of a specialized token or unique identifier, enabling the fine-tuning of the model's embedding based on the provided target concept images or segments of the diffusion model. By em-bedding the representation of the concept into this unique identifier, the model can generate images that are highly consistent with the specified subject or theme.

Several techniques, including textual inversion, ELITE, Custom Diffusion, Dream- Booth, E4T, Imagic, InstantBooth, FastComposer, and Blip-diffusion, have been de- veloped to efficiently and effectively achieve subject-driven image generation. These methods collectively contribute to advancing the capabilities of image generation models and enabling users to create highly personalized and contextually relevant images.

In pursuit of advancing text-conditioned image generation, the paper [20] proposes the kNN-Diffusion method. This approach leverages textual information to retrieve the top k nearest neighbor images from a specified database. These retrieved images then serve as conditional inputs during both the training and inference stages, pro- viding a structured framework for contextualized image generation. Furthermore, the paper introduces the Re-imagen approach, marking a notable departure from conventional methods. It creatively conditions image generation on a dual input, combining retrieved textual descriptions and corresponding images [21]. This dual- conditioning strategy enhances the model's ability to integrate additional context and information, thereby facilitating more nuanced and context-aware image syn- thesis. Collectively, the paper conducts a comprehensive exploration of methods for image generation with text conditioning, encompassing CLIP, RDM, kNN-Diffusion, and Re-imagen. This effort represents a significant stride toward enriching the do-main of image synthesis with textual context, offering promising avenues for future research and practical applications in the field.

In the above-mentioned survey, the research was grounded in the steady-state diffusion model, where diffusion Probabilistic Models (DDPMs) hinge on two funda- mental processes: the forward process and the reverse process. The forward process involves the generation of a sequence of noisy images from a clean image. In con- trast, the reverse process aims to map these noisy images back to the original clean image. In the

forward process, a sequence of noisy images is generated iteratively by introducing Gaussian noise to each subsequent image in the sequence. The mag- nitude of the noise added to each image is regulated by a diffusion coefficient that varies over time. The Markov property of the forward process ensures that the probability distribution of the next state only relies on the current state and the current diffusion coefficient, bypassing the need to consider the entire history of the process. This inherent property enhances the computational efficiency of simulating and analyzing the behavior of the diffusion process over time.

Conversely, the reverse process deals with a sequence of noisy images, endeavoring to map them back to the clean image by iteratively removing the added noise. This process, often referred to as denoising, is notably more intricate than the forward process, primarily due to the non-linear and ill-posed nature of noise removal. Fig- ure 2 shows general architecture of forward and backward process of the diffusion models. DDPMs address this challenge by training a deep neural network to learn the inverse mapping from noisy images to the clean image. This network is trained using data generated by the forward process. The interplay between the forward and reverse processes forms a pivotal foundation for DDPMs, providing a proba- bilistic framework to generate sequences of noisy images from a clean image and to subsequently map these noisy images back to the original clean image.

For those questioning the rationale behind the stepwise approach adopted in [11] instead of a direct transition from noise to a clear image, it is imperative to consider the insights articulated by the authors of [22]. Their findings emphasize that an instantaneous transition would be infeasible and unlikely to yield desirable outcomes. The gradual progression advocated by these authors not only enhances tractability but ultimately results in superior outcomes.
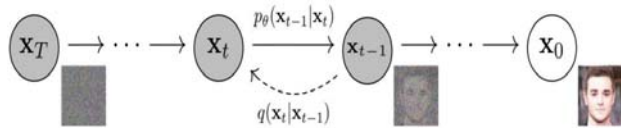

Fig. 2. diffusion model forward and backward process adopted from[18]

Regarding the striving of fostering ethical considerations within the field of artifi-cial intelligence, numerous scholars have endeavoured to address the issue of "bias," a pivotal tenet in the realm of artificial intelligence. In particularly, pertinent in the domain of

generative models due to the potential for biased decisions and the generation of undesirable content. The work presented in [23] discusses the biases in- herent in prevalent generative language models, focusing notably on GPT-2, when applied "out of the box" to downstream tasks. It conscientiously examines biases pertaining to affiliations of various marginalized groups by intersecting gender with dimensions such as religion, ethnicity, political alignment, and geographical origin of names.

The analysis reveals that the machine-predicted jobs are less diverse and more stereotypical for women than for men, especially for intersections. The paper raises the normative question of what language models should learn - whether they should reflect or correct for existing inequalities. The authors suggest that if such models are going to be made readily available, a greater discussion of their fairness and biasis required across more diverse intersectional associations.

In contrast, this paper [24] investigates biases encountered within the trajectory of artistic AI from an art historical perspective. The sociocultural ramifications of these biases are deliberated upon, with the authors aspiring to contribute to the ethical dimensions of generative art. The paper also discusses biases in datasets, specifically representational bias, which arises from having a dataset that is not representative of the real world. The authors mention that this bias can be attributed to multiple factors, including the accessibility and inclusivity of artworks encompassed within the dataset.

Similarly, in the context of art and image generation, this paper [25] introduces a novel methodology to assess social biases in text-to-image systems by generated im- ages across various sociocultural traits. An analysis of over 96,000 images produced by three prominent text-to-image systems reveals a significant overrepresentation of whiteness and masculinity across the systems. Notably, Dall-E 2 is found to be the least versatile, followed by the Stable Diffusion v2 and version 1.4.

These findings underscore the urgency of mitigating 'biases' within text-to-image systems to avert discriminatory outcomes. The paper underscores the significance of delineating the social biases inherent in these systems, bearing in mind their synthetic output and potential adverse consequences.

In addition, in the context of text-to-image generation, this paper [26] introduces novel strategy termed "fair spread" to alleviate biases within generative text-to-image models. These authors reveal that generative AI models, reliant on extensive internet-

derived datasets, can suffer from compromised and prejudiced human be-haviors, perpetuating such biases. The paper expounds upon the adoption of eq-uitable distribution to recalibrate biases across identity groups in accordance with human guidance, obviating the need for data curation or supplementary training. This approach is evaluated using the Stable Diffusion 1.5 model, the LAION-5B dataset (full description of the LAION dataset will be provided below), and a pre-trained CLIP model as detailed in 3.1.1. Moreover, the paper delves into the presence of gender-based biases in occupational contexts within the CLIP model, as ascer-tained by the iEAT test. The iEAT test (image Embedding Association Test) is a method used to evaluate the bias present in learned representations of images. It tests for statistically significant associations between sets of representations, such as encoded images, and attribute sets. The test involves comparing target images of different attributes (e.g., male-appearing people and female-appearing people) against images related to different concepts (e.g., career and family).

The goal is to determine if a biased model associates certain attributes more closely with certain concepts. The test statistic is computed to measure the strength of these associations. The iEAT test has been applied to the CLIP encoder to un- cover gender occupation biases in the encoded images The results indicate that images depicting males tend to be more closely associated with careers, science, and engineering, while images portraying females lean towards familial, artistic, and caregiving domains.

During our investigation, we can conclude that there are great efforts being made to solve the problem of "bias" in generative models, and in particular, text-to-images using the stable diffusion model. However, bridging the gap of generating Saudi culture images is not explored yet. Therefore, in this research we aim to investigate how the model of stable diffusion can generate artistic images that reflect the Saudi culture as a case study and what are possible improvements can be made.

3. Methods

Starting from the bias problem in diffusion generative models and the issue of not generating suitable images that align with Saudi cultural text prompts. The method followed in this work is straightforward and consists of three steps. The first step, incorporating a number of subjects related to Saudi culture. Second step, feed these subjects to the model.

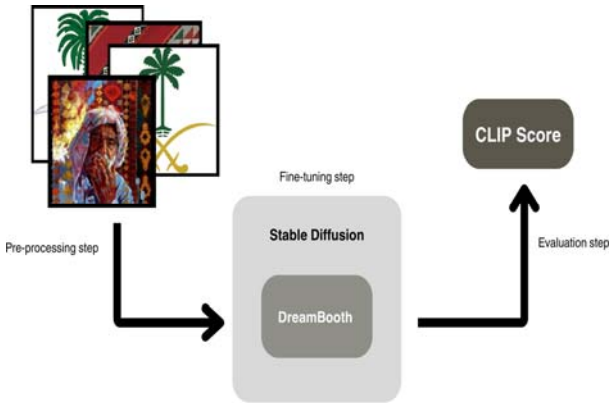Third step, regenerate images with text prompt. Figure3 illustrates the three-stages pipeline of our approach.

Fig. 3. Pipeline of the experiment of injecting Saudi Culture subjects to the model

3.1 The Generative Model

The "Stable Diffusion Model v1-5" is an image generation model built upon the text-to-image diffusion technique. It can create realistic images resembling photographicimages based on textual input. The model utilizes a pre-trained text encoder (CLIPViT-L/14) to generate images and is integrated with the text-to-image diffusion technique during training [13].

3.1.1 CLIP (Text-guided)

CLIP (Counteractive Language-Image Pre-training) is a framework that aligns images and textual descriptions in a joint embedding space. It enables cross-modal understanding by associating images and their corresponding textual descriptions using Counteractive learning. In the Diffusion Model architecture, CLIP is utilizedin a text-guided manner, meaning that textual descriptions are used to guide the generation or modification of images. By leveraging CLIP, the Diffusion Model cangenerate or modify images based on textual input, enabling tasks such as text-to- image synthesis or text-guided image editing [19].

3.1.2 LAION

The model trained on LAION-5B and subsets, which are filtered using LAION's NSFW detector, with a "punsafe" score of 0.1 (conservative) [28]. Datatset columns:

- · URL: the image url,
- ▪ URL: the image url,

- ▪ TEXT: captions, in english,
- ▪ WIDTH: picture width,
- ▪ HEIGHT: picture height,
- ▪ similarity: cosine between text and image ViT-B/32 embeddings, using clip,
- ▪ pwatermark: probability of being a watermarked image,
- ▪ punsafe: probability of being an unsafe image[29].

3.2 Fine-Tuning Methods

In deep learning fine-tuning is a technique where an already pre-trained model is refined or optimized for a new task or dataset. This process enables the extraction and utilization of the valuable information encapsulated within the model's weight parameters, allowing for their effective re-purposing [30].

Numerous fine-tuning methods can be employed with the stable diffusion model, including options like Textual Inversion [31] and DreamBooth. After conducting our own experiment and thorough research, we've determined that DreamBooth out- performs others by being the fastest in the training process and yielding the highest quality outputs. Consequently, we intentionally selected this technique for our ex- periment.

DreamBooth [30] is an advanced fine-tuning technique developed by researchers from Google Research and Boston University in 2022 to refine generative image models through text-based inputs. It relies on a limited dataset to generate novel images of objects, incorporating variations such as changing the subject's location and modifying its properties like color, shape, and viewpoint. We also intend to employ this technique to feed the model and familiarize it with Saudi culture.

The primary task of this technique lies in injecting and embedding the object into the model's output domain to generate new images associated with this object. Itis important to note that fine-tuning the model with a small number of images may lead to mode collapse and overfitting. However, the researchers emphasize that em- ploying a careful fine-tuning setup using diffusion

loss enables the model to excel at integrating new information into its domain without forgetting prior knowledge or overfitting to a small set of training images.

Researchers propose a novel approach that simplifies the need for detailed image descriptions by relying on a unique identifier associated with a broader class name. The class name plays a crucial role in connecting the model's previous knowledge with the specific subject. This methodology minimizes language divergence and en- hances overall performance.

In the pursuit of achieving maximum fidelity in generating images through fine- tuning, the authors identified two significant challenges: language drift and reduced output diversity. Language drift, a phenomenon observed in language models dur- ing specific task fine-tuning, was noted to affect diffusion models, causing them to forget how to generate subjects of the same class. Additionally, fine-tuning on a lim-ited image set risked reducing output diversity, often leading to a convergence on a few-shot perspective. To address these issues, the authors proposed a novel solution, the autogenous class-specific prior preservation loss. This approach counteracts lan- guage drift by supervising the model with its own generated samples, ensuring the retention of prior knowledge. Simultaneously, it enhances output diversity, allowing the model to generate a more varied set of images based on class priors.

3.3 Injection Process Using Dreambooth

Before starting the re-training process, we need to define four standards elements:

1- **TOKEN NAME:** Refers to the unique label that will serve as a reference for the subject being incorporated. This identifier should be distinct to avoid any overlap with existing representations.
2- **CLASS NAME:** This label corresponds to the category name introduced in the initial rationale. While the original DreamBooth paper suggests employing generic categories like "man," "woman," or "child" for human subjects, or "cat" and "dog" for pets.
3- **NUMBER OF REGULARIZATION IMAGES:** the inclusion of class-specific prior-preservation loss is essential to mitigate issues like overfitting and language divergence. We followed the original authors' guidance and included 200 images pertraining image for this

purpose. It's important to note that employing a greater number of regularization images might lead to enhanced outcomes.
4- **TRAINING ITERATIONS:** This parameter denotes the number of iterations executed by the model during the fine-tuning process. If this count is excessively low, the model will inadequately learn the nuances of the subject's images, leading to inaccuracies during inference. Conversely, an excessively high count might result in overfitting, hindering the model's capacity to reproduce the subject in different expressions, poses, or contexts beyond those in the training dataset.

4. Experiments

In this section, we will present our experiment involving the fine-tuning of the model using the DreamBooth technique to enhance the generated images associated with Saudi culture. We will examine the extent of the impact of certain variables, such as the learning rate, and the quantity of images in the model's outputs. Subse- quently, we will showcase the results obtained from this experimentation and the corresponding refinements achieved.

4.1 Image Embedding Association Test (iEAT)

The Image Embedding Association Test (iEAT) quantifies social biases in image embeddings based on semantic similarities [32], designed to discern and quantify social biases that may be embedded within image embeddings. This pioneering method leverages semantic similarities to unearth latent biases within these embeddings.

A significant illustrative application of iEAT is presented in the related work sec- tion above, where it effectively exposes gender-based biases within the CLIP model.The results of the iEAT tests conducted in this context yielded striking insights.It was observed that images depicting males exhibit a higher degree of association with professions, particularly those in the domains of science and engineering. Con- versely, images portraying females were found to be more closely associated with fields pertaining to family, the arts, and caregiving. These results serve as an in- structive illustration of the capability of iEAT to highlight and quantify such biases in image embeddings.

The implications of these findings are vital for addressing and mitigating bias in artificial intelligence

systems and promoting fairness in generative image-related applications. it formulates a test statistic that compares target concepts X and Y with the attributes A and B, defined as:

$$s(X,\ Y,\ A,\ B) = \sum_{x \in X} s(x,\ A,\ B) - \sum_{y \in Y} s(y,\ A,\ B) \qquad (1)$$

where $s(w, A, B)$ is the differential association of w with the attributes, quan-tified by the cosine similarity of vectors

$$s(w,\ A,\ B) = \mu(\cos(w, a)_{a \in A}) - \mu(\cos(w, b)_{b \in B})$$

The statistical significance is determined using a permutation test, contrasting the score $s(X, Y, A, B)$ with the scores $s(Xi, Y i, A, B)$, where Xi and $Y i$ are allequal-sized partitions of the set $X \cup Y$:

$$p_t = Pr[s(X_i, Y_i, A, B) > s(X, Y, A, B)]$$

The effect size d quantifies the bias magnitude, computed as the normalized separation of the association distributions:

$$d = \frac{\mu(s(x,\ A,\ B)_{x \in X}) - \mu(s(y,\ A,\ B)_{y \in Y})}{\sigma(s(t,\ A,\ B)_{t \in X \cup Y})}$$

Baseline: In our research, we employed the 'iEAT' test as a baseline, drawing from the methodology outlined in this paper [26]. This test served as a foundational tool for assessing potential biases in the model's output.

Furthermore, we conducted a comprehensive evaluation to ascertain the presenceof cultural and societal biases, acknowledging that the Stable diffusion model pri- marily relies on descriptions in English. This linguistic limitation can significantly influence the model's outputs, as white and Western cultures are often default reference points.

For our analysis, we specifically examined the term "Sadu". The accompanying Figure 4 demonstrates the outcomes of our investigation. In the top row, we observethe output of the untrained model, which notably lacks recognition of Saudi culture.For instance, prior to model training, the term "Sadu" did not yield any referencesto Saudi Sadu art; instead, it erroneously associated it with a sacred person in the Hindu religion, also known as a "Sadu."

Conversely, in the bottom row, we present the results after retraining the model.Post-training, the model exhibited a more accurate understanding of the term "Sadu," correctly associating it with "Sadu" images, indicative of its enhanced cultural recognition. This exemplifies the model's capacity for adaptability and im- proved cultural awareness through training and highlights the significance of mit-igating cultural biases in AI systems. The output of the stable diffusion model is showed in Figure4, (top row) before training the model does not show the Saudi Sadu art, but rather refers to a sacred person in the Hindu religion also called "Sadu". In contrast, the test result (bottom row) after training the model shows the "Sadu" images after training the model.

Fig. 4. The output of the stable diffusion model before and after training the model with "Sadu" cultural symbol

4.2 Dataset and Evaluation

Dataset: In this particular experiment, we have undertaken the collection of more than 500 images of Saudi cultural symbols and around 30 images pertaining to the art of "Sadu." The Sadu is a traditional form of intricate weaving and embroidery deeply rooted in the cultural heritage of the Kingdom of Saudi Arabia, particularly prevalent in the northern region. The dataset is uploaded on Kaggle[a].

Despite the model's prior training on a dataset encompassing such images, we encountered a notable impediment when attempting to generate images representa-tive of "Sadu" art through command prompts. The model consistently failed to do so prompting us to investigate the underlying reasons for this deficiency. Whereas inthe Indian language, the term 'Sadu' refers to a religious man, completely opposite to the meaning of this term in the Arabic language, where it denotes an artistic pattern. This is what led to the model's poor outputs when asked to generate thingsrelated to 'Sadu'.

Consequently, we made the strategic decision to amass a curated dataset specificallyfocused on "Sadu" art. This dataset serves the dual purpose of reintroducing the concept to the model, along with associating it with a unique identifier. This iden- tifier facilitates rapid and accurate access to the object in question. Additionally, we have assigned a dedicated class name to this object, which enables the model to leverage its previous knowledge and experience for enhancing the quality of its output.

Evaluation Metric: The evaluation of the model's performance was conductedafter training it with "Sadu" images, using the CLIP Score Metric. This metric quantifies the agreement between pairs of image captions, where higher CLIP scores reflect stronger agreement. The CLIP score serves as a quantitative measure of the qualitative concept of "compatibility" and can be seen as a representation of the semantic similarity between an image and its caption. Notably, the CLIP score demonstrates a high correlation with human judgment, as presented in this paper [33][34]. defined as:

$$CLIPScore(I, C) = max(100 * cos(Ei, Ec), 0) \qquad (2)$$

which corresponds to the cosine similarity between visual CLIP embedding Ei foran image i and textual CLIP embedding Ec for an caption c.

The score in this metric ranges from 0 to 100, with higher values indicatingbetter agreement between image and caption. However, as indicated in the table provided 1 2, the results of the overall Score Metric evaluation for our experiments ranged between 22 and 34. These results are unsatisfactory, particularly given

that this evaluation pertains to human judgment.

It is important to recognize that the CLIP Score

Metric primarily assesses the com- patibility of the prompt with the image, and as such, may not offer a comprehensive evaluation of the overall quality of the experiment. Therefore, the inclusion of ad- ditional metrics, such as image similarity, may be imperative to provide a more holistic understanding of the model's performance.

4.3 Results

In fine-tuning experimentation with the model, we utilized diverse datasets compris-ing different quantities of images—5, 10, 15, and 20. Simultaneously, we conducted a study on varying the learning rate, both high

and low, to assess the model's post-retraining performance. This study incorporated a dataset focused on Saudi Arabian cultural symbols, denoted as "Sadu."

Before retraining and implementing the DreamBooth technique, the model's out- puts deviated from Saudi

Arabian cultural symbols, as depicted in Figure 5. These outputs were generated using the prompt text: Sadu pattern wrapping a tent in the desert.

Fig. 5.   Images generated from the Stable Diffusion Model V1-5 before fine-tuning do not reflectthe true meaning of "Sadu."

This discrepancy underscored the need for further fine-tuning and adjustmentto align the model's outputs with the specific nuances of Saudi Arabian culture and customs.

Tables 1 and 2 present the results of model training at high and low learning rates. The higher learning rate demonstrated superior output quality for the "sadu" class, despite some imperfections post fine-tuning with the DreamBooth technique. Theseimperfections were linked to inherent issues in the diffusion model.

In some outputs, typing "Sadu" in the command prompt displayed an image of an elderly man with a Hindu appearance, later identified as a religious figure known as a "Sadhu." This cultural reference issue stemmed from differences in language and cultural references, particularly in English. Experiments

highlighted the impact of mentioning or omitting the class name in the command prompt. The table revealed that mentioning the class name "pattern" did not generate results related to the command prompt.

A limiting factor affecting output quality is the constraint inthe stylized diffusion model, particularly in handling more complex tasks involving compositionality [13]. This includes scenarios like rendering an image correspondingto a "red cube on top of a blue sphere" or covering objects like a tent with a "Sadu" pattern. The model's image generation performance is influenced by factors such aslearning rate, cultural references, and the limitations of the stylized diffusion model.These aspects necessitate further investigation and refinement for optimal results.

Table 1.   Outputs after model retraining with a **high learning rate of 2e-6** for various numbers of images

Prompt	No. Training Output Images		CLIP SCORE
sadu pattern wrapping a tent in desert	5		30.6932
sadu pattern wrapping a tent in desert	10		29.9615
sadu pattern wrapping a tent in desert	15		26.3822
sadu pattern wrapping a tent in desert	20		27.7432
sadu wrapping a tent in desert	5		24.6013
sadu wrapping a tent in desert	10		22.4935
sadu wrapping a tent in desert	15		22.4863
sadu wrapping a tent in desert	20		23.9169

Post this experiment, the primary objective was to identify the optimal approach for achieving ideal results, specifically in fine-tuning the model using the Dream-Booth technique. The comprehensive analysis focused on the learning rate, various quantities of images used during training, and the inclusion of the class name in the prompt as the results were significantly better in not including it, and the disparity in the number of images did not yield a significant difference. However, a discernible trend underscored the superiority of training objects with a higher learning rate for optimal results.

Table 2.   Outputs after model retraining with a **low learning rate of 1e-6** for various numbers of images

Prompt	No. Training Output Images		CLIP SCORE
sadu pattern wrapping a tent in desert	5		30.4857
sadu pattern wrapping a tent in desert	10		25.9006
sadu pattern wrapping a tent in desert	15		28.1366
sadu pattern wrapping a tent in desert	20		29.7103
sadu wrapping a tent in desert	5		23.5857
sadu wrapping a tent in desert	10		33.4855
sadu wrapping a tent in desert	15		34.8456
sadu wrapping a tent in desert	20		29.9831

To thoroughly validate these findings by specifically examining the model's ca- pacity to recognize additional cultural symbols. Another symbolic is used which is the "Makkah Clock" located within the city of Makkah. As expected, the initial results were unsatisfactory and did not align with the intended cultural reference, as depicted in Figure 6 on the left. However, through meticulous fine-tuning and injecting the model with images of the original tower at a higher learning rate, the outcomes successfully aligned with the intended cultural symbol associated with the "Makkah Clock" on the right

in Figure 6. Results obtained from the pre-trained Stable Diffusion v1-5 model are showcased. On the left, an attempt was made to generate an image of a "Makkah Clock" using the prompt "Oil painting of a Makkah Clock." On the right, the model's outcomes for images representing the "Makkah Clock" are presented after fine-tuning using the DreamBooth technique. These re- fined results capture the "Makkah clock" cultural symbol.

Fig. 6.   Validation results using new cultural symbol "Makkah Clock"

Conversely, in the interest of fairness, when evaluating the model's response to the prompt "pilgrims in Makkah," as illustrated in Figure 7, the generated results were deemed acceptable. However, as people of the culture, we acknowledge that its representation is suboptimal. This aspect highlights the nuanced performance of the model, where certain cultural references were successfully interpreted and reflected in the generated outputs. Another experiment was conducted for further well-known religious symbol "pil-

grims". The results presented in Figure 7. On the left, an attempt was made to generate an image of a "pilgrims" using the prompt: "Oil painting of pilgrims in Makkah." Despite understanding the prompt correctly, the results were sub-optimal in capturing cultural symbols. On the right, the model's outputs refined images of "pilgrims" engaging in their sacred rituals at the holy sites in Makkah after utilizing our approach integrated with DreamBooth technique.

Fig. 7.   example of the prompt "Oil painting of pilgrims in Makkah"

Conversely, in the interest of fairness, when evaluating the model's response tothe prompt "pilgrims in Makkah," as illustrated in Figure 7, the generated results were deemed acceptable. However, as people of the

5. Discussion and Conclusion

In our experiments, we undertook parameter fine-tuning associated with Dream- Booth technology[35], with the overarching objective of generating outputs that faith-fully encapsulate the essence of Saudi culture. The primary aim was to rectify biasesagainst Saudi culture by integrating images representative of its various facets into the model. The objective was to produce images that authentically capture the identity of Saudi Arabia. Our experimental findings attest to the efficacy of Dream- Booth in terms of output quality and training speed. However, challenges emerged in dealing with synthetic images, particularly in the context of superimposing "Sadu" reliefs onto a tent, as delineated in Table 1 and 2. The model exhibited difficulty in preserving the authenticity of the original image, as evidenced by the results. Table 2, focusing on training with lower learning rates, exhibited higher evaluation scores using the CLIP measure. It is noteworthy that the table presenting results from training with a higher education rate successfully captured "Sadu" inscriptions as- sociated with Saudi culture. This observation reinforces our contention that biases originate from a deficiency in Arabic descriptions aligned with Saudi cultural sym- bols within the dataset. Importantly, discrepancies arose when the model assigneda higher rating to a holy man named "Sadu" while assigning lower ratings when "Sadu" inscriptions were clear.

The model also struggled to accurately depict the "Makkah Clock" in its at-tempt, yet it adequately represented "the pilgrims." However, this does not negate our characterization of bias, given our confidence in the existence of concealed sym- bols that prove challenging to represent.

Our study aims to facilitate access to Saudi culture within generative models. Although existing datasets used for model training incorporate images depicting cultural facets, their captions often pose challenges due to non-human curation efforts[28]. Therefore, our research focuses on training the model using images that faithfully portray Saudi culture, accompanied by precise captions to enhance data accessibility.

In conclusion, our experimental methodology offers a resolution to the identified problem. However, acknowledging our profound cultural connection and dedication to human evaluation, we recognize the imperative of continuous improvement in these outcomes.

In our experiments of non-creative Islamic symbols, the model encountered chal- lenges in accurate representation. To enhance the stable diffusion model, we proposean effective strategy involving the augmentation of existing large-scale image cap- tion datasets through re-captioning. This encompasses training and fine-tuning the CLIP image caption model to generate concise or detailed synthetic captions. Our findings suggest that training text-to-image models primarily on detailed synthetic captions yields optimal results. Additionally, the integration of large language mod- els (LLMs), such as OpenAI's DALLE 3, and leveraging advancements in GPT-4, has proven effective in achieving desired outcomes[36]. Notably, Google preceded them and has also implemented a similar methodology in their diffusion model Imagen.[14].

Furthermore, we underscore the importance of implementing metrics specifically designed to evaluate subject accuracy, including the preservation of subject details inthe generated images. Metrics such as DINO, as detailed in a relevant paper[30], offer a promising avenue to achieve this objective. However, due to certain limitations, our access to DINO and its implementation is constrained. Therefore, future workin this domain should explore opportunities to incorporate multiple metrics for a more comprehensive evaluation of generated images.

Acknowledgments

References

[1] D. E. Rumelhart, G. E. Hinton and R. J. Williams, Learning representations by back-propagating errors, *Nature* **323** (1986) 533–536.

[2] H. L. Jungang Xu and S. Zhou, An overview of deep generative models, *IETE Tech-nical Review* **32**(2) (2015) 131–139.

[3] A. Razavi, A. Van den Oord and O. Vinyals, Generating diverse high-fidelity imageswith vq-vae-2, *Advances in neural information processing systems* **32** (2019).

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G.

Sastry, A. Askell *et al.*, Language models are few-shot learners, *Advances in neural information processing systems* **33** (2020) 1877–1901.

[5] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford and I. Sutskever, Jukebox: A generative model for music, *arXiv preprint arXiv:2005.00341* (2020).

[6] M. Kumar, M. Babaeizadeh, D. Erhan, C. Finn, S. Levine, L. Dinh and D. Kingma, Videoflow: A flow-based generative model for video, *arXiv preprint arXiv:1903.01434***2**(5) (2019) p. 3.

[7] T. Marwah, G. Mittal and V. N. Balasubramanian, Attentive semantic video genera-tion using captions, in *Proceedings of the IEEE international conference on computer vision*2017, pp. 1426–1434.

[8] E. I. Nikolaev, Opportunities and challenges in deep generative models, in *CEUR Workshop Proceedings*2018, pp. 326–329.

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

[10] M. Elasri, O. Elharrouss, S. Al-ma'adeed and H. Tairi, Image generation: A review, *Neural Processing Letters* **54** (03 2022).

[11] J. Ho, A. Jain and P. Abbeel, Denoising diffusion probabilistic models, *Advances in neural information processing systems* **33** (2020) 6840–6851.

[12] P. Dhariwal and A. Nichol, Diffusion models beat gans on image synthesis, *Advances in neural information processing systems* **34** (2021) 8780–8794.

[13] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, High-resolution image synthesis with latent diffusion models, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*June 2022, pp. 10684–10695.