Text-to-Image Generation for Art and Society

Submitted to

Graduate School of Information Science and Technology

The University of Osaka

July, 2025

Yankun WU

Abstract

Text-to-image generation models have attracted increasing attention due to their remarkable capabilities in producing high-quality images given natural language prompts. From a positive perspective, these models provide novel opportunities to address challenges that were difficult to solve. However, they also inherit and amplify societal biases, reproducing stereotypes related to gender, race, and age in the generated images.

To examine both their potential for downstream applications and their ethical risks, this thesis investigates the impact of text-to-image generation models on both art and society. First, we explore how generative models can be used to improve digital art analysis. We propose GOYA, which leverages synthetic images generated by Stable Diffusion and employs contrastive learning to disentangle content and style in art paintings. GOYA demonstrates strong performance in downstream tasks such as classification and retrieval, highlighting the potential of generative models in the digital humanities.

Despite their promising applications in representation learning, the inherent bias of generative models cannot to overlooked. To address these ethical concerns, we introduce an automatic protocol to evaluate gender bias in text-to-image generation. Through a systematic analysis of representational disparities, object co-occurrence similarity, and prompt-image dependencies, we reveal that gender bias originates from text embeddings, propagates through the generation process, and is reflected in the entire output image. Based on these findings, we present a training-free method for mitigating gender bias. Our approach interpolates between feminine and masculine embeddings within both the text embeddings and the attention module of the model to generate fairer and diverse neutral outputs. This method does not require model fine-

tuning or additional data, offering a lightweight and practical solution for improving fairness in image generation.

With these contributions, this thesis offers a comprehensive view of the capabilities and risks of text-to-image models, shedding light on their potential to enhance representation learning while underscoring the importance of addressing their social impact.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recent breakthroughs in text-to-image generation have greatly advanced the ability to translate natural language into high-quality visual content. Pioneering models such as Stable Diffusion [1, 3] and DALL-E 2 [4] exhibit remarkable capabilities in producing high-fidelity images semantically aligned with the natural image prompts. These developments have opened up new possibilities across a wide range of fields, including design, entertainment, and digital art. As generative models become increasingly integrated into daily life, discussions have emerged regarding their impact on artistic representation and ethical considerations. This thesis investigates how text-to-image diffusion models affect both art and society, which motivates our major focus: **how does text-to-image generation affect art and society, and how can we enhance its benefits while minimizing its harms?**

In addition to image generation, text-to-image models also serve as powerful tools across various computer vision tasks, such as object detection [5], segmentation [6], and representation learning [7]. The representations of the generated images provide rich information that can be used for downstream tasks [7]. From an artistic perspective, these models offer novel opportunities for exploring visual semantics and aesthetics. In particular, by leveraging these representations, we can achieve a better understanding of concepts such as content and style in artwork, providing new pathways for the use of generative models in digital humanities beyond image generation.

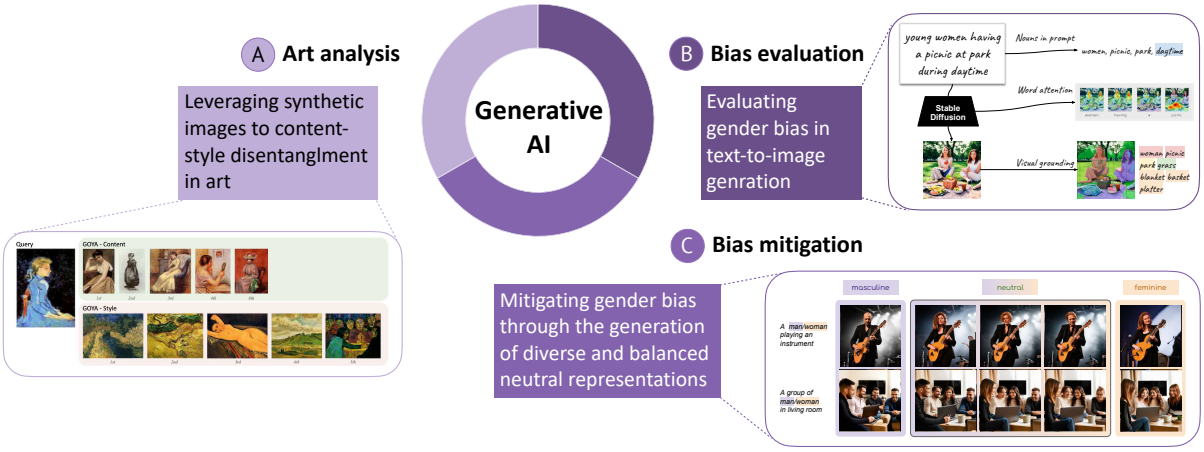

Figure 1.1: Overview of the three projects in this thesis.

However, on the other side, their ethical implications for society cannot be overlooked. These models often contain or even amplify societal bias embedded in the generated images [8–11]. Prior studies have shown that the generated images may reinforce harmful stereotypes related to gender [8–10], race [9, 10], age [11], etc. By investigating the internal components of generative models, we examine how bias emerges throughout the generation process, and propose practical strategies for bias mitigation.

This thesis consists of three projects, each addressing a different aspect of how text-to-image generation is applied to art analysis and its impact on society (Figure 1.1).

In Chapter 2, we explore how generative models can contribute to digital art analysis. Specifically, we address the challenge of disentangling *content* and *style* in paintings – two fundamental elements yet intertwined elements that jointly shape the visual and semantic attributes of artworks. While prior work in digital art analysis often relies on supervised classification using categorical labels (*e.g.*, artist, style, etc), such methods fail to capture the subtle semantic (*content*) and visual (*style*) features present in individual artworks. To solve this issue, we propose GOYA, a novel framework that leverages synthetic images generated by Stable Diffusion from designed prompts that contain both content and style descriptions. Using contrastive learning, GOYA disentangles these two components without the need for human anno-

tations. Evaluation on the WikiArt dataset [2] demonstrates that the disentangled embeddings have good performance on classification and image retrieval while achieving effective disentanglement between content and style. This chapter positions generative models as a powerful tool for representation learning in the digital humanities.

On the society side, Chapter 3 focuses on the ethical considerations of generative models, specifically evaluating gender bias in text-to-image generation. Although prior research has examined demographic bias in facial attributes and on occupation-based prompts, less attention has been paid to how bias originates and propagates during the generation process. To address this, we develop an automatic evaluation protocol for evaluating gender bias in Stable Diffusion. We construct neutral, feminine, and masculine as prompt triplets derived from captions in vision-language datasets and one sentence set generated by ChatGPT [12]. The resulting images are evaluated across three spaces: 1) representational similarity in several spaces during generation, 2) object co-occurrence in the output images, and 3) prompt-image dependency that reflects the alignment between the textual inputs and the generated visual content. Our findings reveal a consistent pattern that neutral prompts tend to produce outputs that are more visually and semantically aligned with those generated from masculine prompts than from feminine ones. Moreover, object-level analysis further shows that the concepts such as clothing, surroundings, and background have unbalanced correlations with the gendered prompts. This chapter contributes a comprehensive and automatic evaluation protocol for evaluating bias beyond facial features, revealing that bias originates in the text embeddings and manifests throughout the entire images.

Based on these findings, Chapter 4 presents an efficient strategy for mitigating gender bias without model fine-tuning or extra data. We introduce a training-free interpolation framework that manipulates the internal representations within the latent space of a pre-trained text-to-image model, producing fairer and diverse neutral outputs. Given a neutral prompt, we first construct its feminine and masculine counterparts and then interpolate between their embeddings to construct semantically neutral representations. The interpolation is applied both in the text embedding and attention module of Stable Diffusion 3 [1], allowing for real-time modulat-

ing of the generation process. To enhance coherence, we further employ a Beta distribution to dynamically sample interpolations and select the most perceptually consistent path. Our experiments show that this method effectively reduces representational disparities between gendered prompts while maintaining high image quality and diversity. As a plug-and-play approach, this method requires no additional data or retraining, offering a lightweight solution for bias mitigation in real-world deployments.

Overall, these three projects contribute both methodological innovations and analytical insights into understanding the advantages and limitations of generative models. By investigating both artistic representations and social bias, we provide a multi-dimensional exploration into the impact of text-to-image models on art and society. The main contributions are as follows:

- A novel approach that leverages generative models to disentangle content and style in artworks, demonstrating the potential of generation models beyond generation but in digital humanity.
- A systematic evaluation protocol for evaluating gender bias in diffusion generative models, revealing how bias manifests beyond facial attributes to the entire image.
- A training-free bias mitigation method that manipulates the latent embeddings within the text and attention modules in the text-to-image model, yielding balanced and diverse outputs for neutral without altering model weights.

By critically engaging with both the strengths and risks of generative models, this work aims to guide future research and promote the ethical deployment of AI systems in creative and socially impactful domains.

Chapter 2

Leveraging Generative Art for Content-Style Disentanglement

2.1 Overview

Content and style are two fundamental elements in the analysis of art. Content refers to the subject matter depicted in the artwork, answering the question of *what scene the artwork depicts*, e.g., a girl chasing a butterfly, fruits on a table, or a street scene near a river. On the other hand, style corresponds to *how the artwork looks*, focusing on the visual appearance of the image, such as color compositions, brushstrokes, and perspective. Each artwork is characterized by a distinctive integration of content and style, making the disentanglement of these two elements an essential aspect of the study of digital humanities.

While humans can easily distinguish content and style, from a computer vision perspective, the boundary between content and style is not so clear. Generally, in the computer vision field, object detection techniques are widely applied to analyzing content in artworks [13]. However, artworks may contain similar objects while still conveying different *subject matters*. Similarly, the automatic analysis of style presents its own challenges. Without a formal definition of what visual appearance is, there is a degree of vagueness and subjectivity in the computation of style.

Some methods [14, 15] classify style by relying on well-established attributes, such as author or artistic movement. While this approach may work on certain applications, such as artist identification [16], it may not be applicable to other tasks such as style transfer [17] or image search [18]. In style transfer, for example, style is defined as the low-level features of an image (e.g., colors, brushstrokes, shapes). However, in a broader sense, style is not formed by a single image but by a set of artworks that share a common visual appearance [19].

To address these challenges, most methods for art analysis rely on full supervision [15, 20], requiring corresponding content or style labels for each image in the dataset. Although some art datasets with labeled attributes are available (e.g., WikiArt [2], The Met [18], APOLO [21]), additional issues arise. Firstly, the attributes of new artworks still require experts to annotate them. Moreover, the annotated labels commonly are words describing general traits of artwork collections, making it difficult to convey subtle differences between artworks. For instance, what scene does a painting in the *still life* genre depict? What does the visual appearance of an *Expressionism* style painting look like? While we can infer some of the common attributes they may carry, e.g., inanimate subjects in the *still life* painting and strong subjective emotions in the *Expressionism* painting, detailed attributes such as depicted concepts, color composition, and brushstrokes still remain unknown. When training based on labels, it is challenging to capture the subtle content and style discrepancies in images. To resolve this problem, some work [22] leverages natural language descriptions instead of categorical classes. Although natural language can overcome the ambiguity and rigidity of labels, they still require human experts to write descriptions for each image.

In our work, we exploit the generative power of a popular text-to-image model, Stable Diffusion [3], and propose leveraging the distilled knowledge as a prior to learn disentangled content and style embeddings of paintings [23, 24]. Given a prompt specifying the desired content and style, Stable Diffusion can generate a diverse set of synthetic images while maintaining consistency with the prompt. The subtle characteristics of content and style in the synthetically generated images can be controlled through well-defined prompts. Thus, free from direct human annotations, we train on the generated images to disentangle content and style using

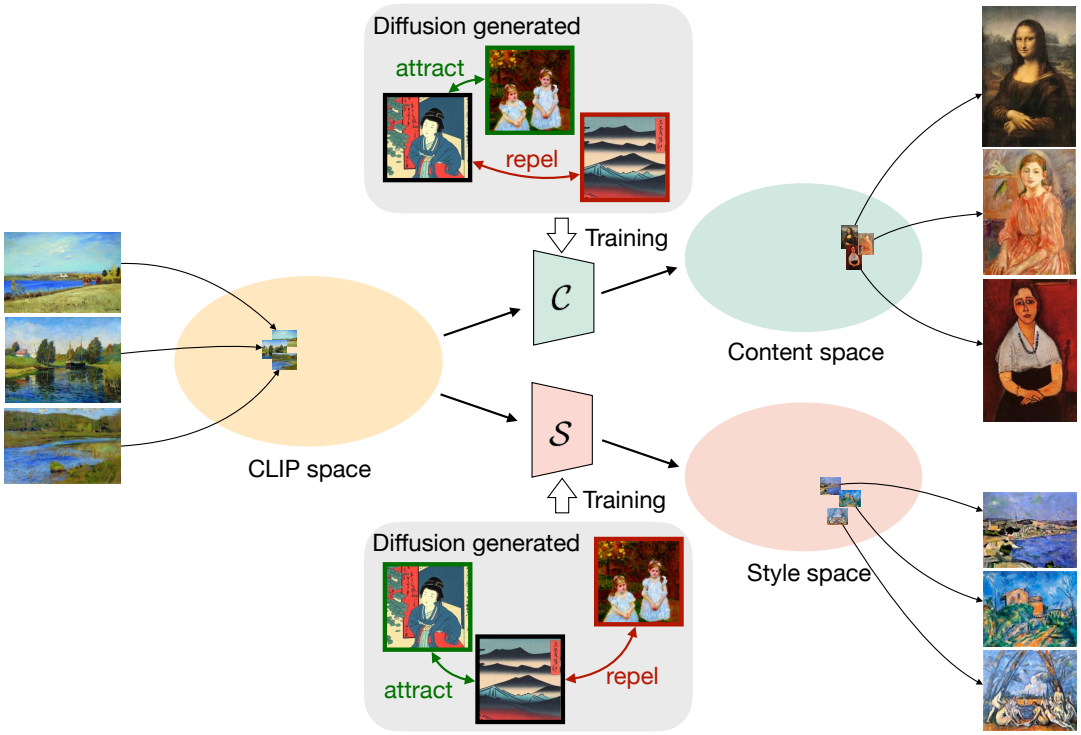

Figure 2.1: An overview of our method, GOYA. By using Stable Diffusion generated images, we disentangle content and style spaces from CLIP space, where content space represents semantic concepts and style space captures visual appearance.

contrastive learning. Previous work also shows that Stable Diffusion generated images can be useful for image classification [25].

The intuition behind our method, named GOYA (disentanGlement of cOntent and stYle with generAtions), is that, although there is no explicit boundary between different contents or styles, significant dissimilarities can be distinguished by comparison. Our simple yet effective model (Figure 2.1) first extracts joint content-style embeddings using a pre-trained Contrastive Language-Image Pretraining (CLIP) image encoder [26], and then applies two independent transformation networks to learn disentangled content and style embeddings. These transformation networks are trained on the generated synthetic images with contrastive learning, reducing the reliance on human image-level annotations.

We conducted three tasks and an ablation study on a popular benchmark of paintings, the WikiArt dataset [2]. We show that, even with distilled knowledge from Stable Diffusion, our model achieves better disentanglement between content and style compared to other models trained on real paintings. Additionally, experiments demonstrate that the resulting disentangled spaces are useful for downstream tasks such as similarity retrieval and art classification. In summary, our contributions are as follows:

- We design a disentanglement model to obtain disentangled content and style space derived from CLIP's latent space.
- We train our model with synthetic images rather than real paintings, leveraging the capabilities of Stable Diffusion and prompt design.
- Results indicate that the knowledge in Stable Diffusion can be effectively distilled for art analysis, performing well in content-style disentanglement, art retrieval, and art classification.

Our findings pave the way for the adoption of generative models in digital humanities, not only for generation but also for analysis. The code is available at `https://github.com/yankungou/GOYA`. [1]

2.2 Related Work

2.2.1 Art Analysis

The use of computer vision techniques for art analysis has been an active research topic for decades, particularly in tasks such as attribute classification [27,28], object recognition [29,30], and image retrieval [13, 18]. Fully-supervised tasks (e.g., genre or artist classification [27]) have achieved outstanding results by leveraging neural networks trained on annotated datasets [31,32]. However, image annotations have some limitations, particularly in the categorization of

[1] This chapter is based on the conference paper [23].

styles. Multiple datasets [33–36] provide style labels, which abundant research [16, 37–39] has utilized for style classification. This direction of work assumes style to be a static attribute rather than dynamic and evolving [19]. A different interpretation is provided by style transfer [17] where a model extracts the low-level representation of a *stylized image* (e.g., a painting) and applies it to a *content image* (e.g., a plain photograph), defining style based on a single artwork's characteristics like color, shape, and brushstroke. To address the limitations of rigid labels in supervised learning and the narrow focus on a single image in style transfer, we propose learning disentangled embeddings of content and style through similarity comparisons leveraging the flexibility of a text-to-image generative model.

2.2.2 Representation Disentanglement

Disentangling representation plays an essential role in various computer vision tasks such as style transfer [40, 41], image manipulation [42, 43], and image-to-image translation [44, 45]. The goal is to discover discrete factors of variation in data, thus improving the interpretability of representations and enabling a wide range of downstream applications. Previous work on disentangling attributes like azimuth, age, or gender has utilized adversarial learning [46] or variational autoencoders [47], aiming to encourage discrete properties in a single latent space. For content and style disentanglement, approaches apply generative models [40], a diffusion model [48], or an autoencoder architecture with contrastive learning [49]. In the art domain, ALADIN [49] concatenates the adaptive instance normalization (AdaIN) [50] feature into the style encoder to learn style embedding for visual searching. Kotovenko et al. [40] propose fix-point triplet loss and disentanglement loss for performing better style transfer. However, these approaches lack semantic analysis of content embeddings in paintings. Recently, Vision Transformer (ViT)-based models has shown the ability to obtain structure and appearance embeddings [51, 52]. DiffuseIT [48] and Splice [51] learn content and style embeddings by utilizing the keys and the global $[\mathrm{CLS}]$ token of pre-trained DINO [52]. In our work, taking advantage of the generative model, our approach builds a simple framework to decompose the latent space

into content and style spaces with contrastive learning, exploring the use of generated images in representation learning.

2.2.3 Text-to-Image Generation

Text-to-image generation models aim to produce synthetic images based on given text inputs. Fueled by datasets containing vast text-image pairs that have emerged in recent years, numerous text-to-image generation models have been developed [3,4,53]. For instance, CogView [53] is trained on 30 million text-image pairs, while DALL-E 2 [4] is trained on 650 million text-image pairs. One of the main challenges faced by these models is achieving semantic coherence between guiding texts and generated images. This challenge has been addressed by using pre-trained CLIP embeddings [26] to construct aligned text and image features in the latent space [54–56]. Another challenge is obtaining high-resolution synthetic images. GAN-based models [57, 58] have shown good performance in improving the quality of generated images; however, they suffer from instability during training. Leveraging the superior training stability, approaches based on diffusion models [3] have recently emerged as a popular tool for generating near-human quality images. Despite the rapid development of models for image generation, how the features of synthetic images can be utilized remains an underexplored area of research. In this paper, we study the potential of generated images for enhancing representation learning.

2.2.4 Training on Synthetic Images

With the increasing availability of open-sourced applications in generative models, synthetic images can be collected and integrated into training data, potentially impacting the development and performance of future models [59]. Several studies have investigated the impact of synthetic images across various aspects, including art forgeries [60], learnt representations [7], datasets [61], model training [62,63], and classification [25,61]. Tian et al. [7] demonstrate that training solely on synthetic images using self-supervised methods can yield better representations than training on real images of the same sample size. Sariyildiz et al. [25] show that models trained

on synthetic ImageNet clones achieve comparable performance on classification tasks to those trained on real image. Azizi et al. [62] demonstrate that augmenting real data with generated images during training improves classification accuracy score (CAS) [64]. In the art domain, Ostmeyer et al. [60] find that training with synthetic images enhances the recognition of human-made art forgeries. In our work, we explore leveraging synthetic images for content and style disentanglement in art paintings.

2.3 Preliminaries

2.3.1 Stable Diffusion

Diffusion models [3, 65] are generative methods trained in two stages: a forward process with a Markov chain to transform input data to noise, and a reversed process to reconstruct data from the noise, achieving high-quality performance in image generation.

To reduce training costs and accelerate the inference process, Stable Diffusion [3] trains the diffusion process in the latent space instead of the pixel space. Given a text prompt as input condition, the text encoder transforms the prompt to a text embedding. Then, by feeding the embedding into the UNet through a cross-attention mechanism, the reversed diffusion process generates an image embedding in the latent space. Finally, the image embedding is fed to the decoder to generate a synthetic image.

In this work, we define symbols as follows: given a text prompt $x = \{x^C, x^S\}$ as input, we can obtain the generated image y. The text x^C represents content description and x^S denotes style description, where $\{\cdot\}$ indicates a comma-separated string concatenation.

2.3.2 CLIP

CLIP [26] is a text-image matching model that aligns text and image embeddings in the same latent space. It shows high consistency between the visual concepts in the image and the semantic concepts in the corresponding text. The text encoder $\mathcal{E}_T$ and image encoder $\mathcal{E}_I$ of CLIP

are trained with 440 million text-image pairs, showing outstanding performance on various text and image downstream tasks, such as zero-shot prediction [66, 67] and image manipulation [55,56,68]. Given the text x and an image y, the CLIP embeddings f from text, and g from image, both in $\mathbb{R}^d$, can be computed as follows:

$$f = \mathcal{E}_T(x), \tag{2.1}$$

$$g = \mathcal{E}_I(y). \tag{2.2}$$

To exploit the multi-modal CLIP space, we employ the pre-trained CLIP image encoder $\mathcal{E}_I$ to obtain CLIP image embeddings as the prerequisite for the subsequent disentanglement model. Moreover, during the training stage, the CLIP text embedding of a prompt is applied to acquire the semantic concepts of the generated image.

2.4 GOYA

We aim to learn the disentangled content and style embeddings of artworks in two different spaces. To collect a diverse set of artistic images with various content and style, we leverage Stable Diffusion to generate synthetic images based on specific content and style descriptions. By training with contrastive loss, our GOYA model effectively learns the proximity of different artworks in two spaces, guided by text prompts.

Figure 2.2 shows an overview of GOYA. Given a mini-batch of N prompts $\{x_i\}_{i=0}^{N}$, where $x_i = \{x_i^C, x_i^S\}$ with comma-connected content and style descriptions, we obtain diffusion generated images y_i using Stable Diffusion. We then compute CLIP image embeddings g_i by Equation (2.2) and use a content and a *style encoder* to obtain disentangled content and style embeddings in two different spaces, respectively. As previous research has shown [69] that content and style possess different properties, while content embeddings correspond to higher layers in the deep neural network and style embeddings correspond to lower layers. Accordingly, we design an asymmetric network architecture for extracting content and style, a common approach in the art analysis domain [32, 40, 49, 69].

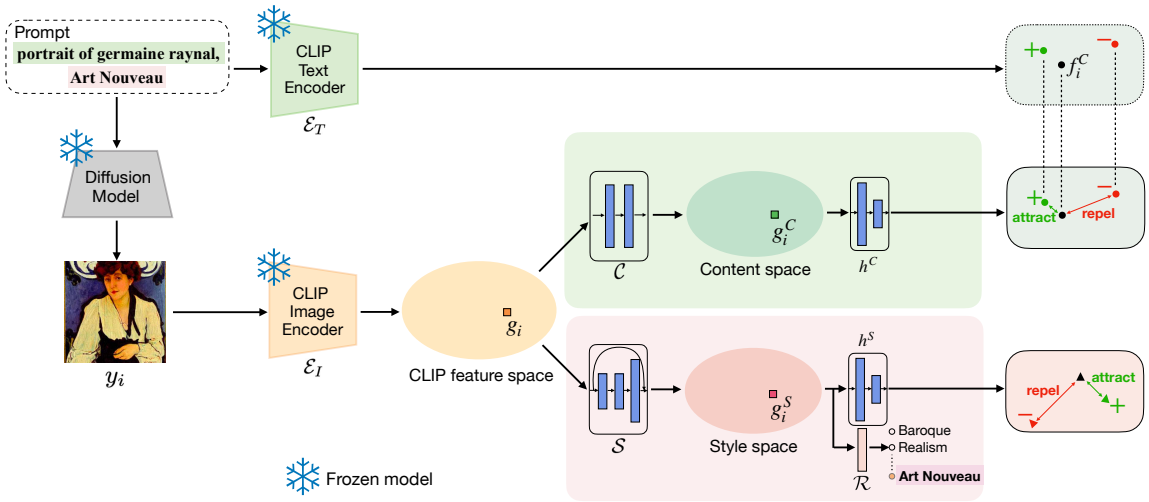

Figure 2.2: Details of our proposed method, GOYA, for content and style disentanglement. Given a synthetic prompt containing content (first part of the prompt, in green) and style (second part of the prompt, in red) descriptions, we generate synthetic diffusion images. We compute CLIP embeddings with the frozen CLIP image encoder, and generate content and style disentangled embeddings with two dedicated encoders $\mathcal{C}$ and $\mathcal{S}$, respectively. In the training stage, projectors h^C and h^S and style classifier $\mathcal{R}$ are used to train GOYA with contrastive learning. For content, contrastive learning pairs are chosen based on the text embedding of content description in the prompt extracted by frozen CLIP text encoder. For style, contrastive learning pairs are chosen based on the style description in the prompt.

2.4.1 Content Encoder

The content encoder $\mathcal{C}$ maps CLIP image embedding g_i to content embedding g_i^C as follows:

$$g_i^C = \mathcal{C}(g_i), \tag{2.3}$$

$\mathcal{C}$ is a two-layer perceptron (MLP) with ReLU non-linearity. Following previous research [70], to make content g_i^C highly linear, during training, we add a non-linear projector h^C on top of the content encoder, which is a three-layer MLP with ReLU non-linearity.

2.4.2 Style Encoder

Style encoder $\mathcal{S}$ also maps CLIP image embedding g_i but to style embedding g_i^S as follows:

$$g_i^S = \mathcal{S}(g_i). \tag{2.4}$$

$\mathcal{S}$ is a three-layer MLP with ReLU non-linearity. In particular, following [71], we apply a skip connection before the last ReLU non-linearity in $\mathcal{S}$. Similar to the content encoder, non-linear projector h^S with the same structure as h^C is added after $\mathcal{S}$ to facilitate contrastive learning.

2.4.3 Content Contrastive Loss

Unlike prior research [40], which defines content similarity only solely based on style-transferred images originating from the same source, we use a broader definition of content similarity. We introduce a soft-positive selection strategy that identifies pairs of images with similar content according to their semantic similarity. That is, two images sharing similar semantic concepts are designated as a positive pair, whereas images lacking semantic similarity are considered negative pairs.

To quantify *semantic similarity* between a pair of images, we exploit the CLIP latent space and conduct text similarity between the associated texts. Given the content description x_i^C of the image y_i, we consider the CLIP text embedding $f_i^C = \mathcal{E}_T(x_i^C)$ as a proxy for the content of y_i. Therefore, for a pair of two diffusion images (y_i, y_j) and a text similarity threshold ϵ^T, they are considered a positive pair if $D_{ij}^T \leq \epsilon^T$, where D_{ij}^T is the text similarity obtained by the cosine distance between the CLIP text embedding f_i^C and f_j^C. The content contrastive loss is defined as follows:

$$L_{ij}^C = \mathbb{1}_{[D_{ij}^T \leq \epsilon^T]}(1 - D_{ij}^C) + \mathbb{1}_{[D_{ij}^T > \epsilon^T]} \max(0, D_{ij}^C - \epsilon_c), \tag{2.5}$$

where $\mathbb{1}_{[\cdot]}$ is the indicator function that yields 1 when the condition is true and 0 otherwise. D_{ij}^C is the cosine distance between $h_C(g_i^C)$ and $h_C(g_j^C)$, which are the content embeddings of images after projection. ϵ_c is the margin that constrains the minimum distance of negative pairs.

2.4.4 Style Contrastive Loss

The style contrastive loss is defined based on the style description x^S given in the input prompt. If a pair of images share the same style class, then they are considered a positive pair, indicating that their style embeddings should be close in the style space. Otherwise, they are deemed a negative pair, and they should be pushed away from each other. Given (y_i, y_j), the style contrastive loss can be computed as follows:

$$L_{ij}^S = \mathbb{1}_{[x_i^S = x_j^S]}(1 - D_{ij}^S)) + \mathbb{1}_{[x_i^S \neq x_j^S]} \max(0, D_{ij}^S - \epsilon^S), \tag{2.6}$$

where D_{ij}^S is the cosine distance between the style embeddings $h^S(g_i^S)$ and $h^S(g_j^S)$ after projection, and ϵ^S is the margin.

2.4.5 Style Classification Loss

To learn the general attributes of each style, we introduce a style classifier $\mathcal{R}$ to predict the style description (given as x_i^S) based on the embedding g_i^S of image y_i. Prediction w_i^S by the classifier is given by

$$w_i^S = \mathcal{R}(g_i^S), \tag{2.7}$$

where $\mathcal{R}$ is a linear layer network. For training, we use softmax cross-entropy loss, which is denoted by L_i^{SC}. Note that the training of this classifier does not rely on human annotations, but on the synthetic prompts and generated images by Stable Diffusion.

2.4.6 Total Loss

In the training process, we compute the sum of three losses. The overall loss function in a mini-batch is formulated as

$$L = \lambda^C \sum_{ij} L_{ij}^C + \lambda^S \sum_{ij} L_{ij}^S + \lambda^{SC} \sum_{i} L_i^{SC}, \tag{2.8}$$

where λ^C, λ^S, and λ^{CS} are parameters to control the contributions of losses. We set $\lambda^C = \lambda^S = \lambda^{CS} = 1$. The summations over i and j are computed for all pairs of images

in the mini-batch, and the summation over i is for all images in the mini-batch.

2.5 Evaluation

We evaluate GOYA on three tasks: disentanglement (Section 2.5.5), classification (Section 2.5.7), and similarity retrieval (Section 2.5.6). We also conduct an ablation study in Section 2.5.8.

2.5.1 Evaluation Data

To assess content and style in the classification task, we utilize genre and style movement labels in art datasets that can serve as substitutes for presenting content and style, even if they do not entirely satisfy our definitions in this paper. In detail, the genre labels indicate the type of scene depicted in the paintings, such as "portrait" or "cityscape", while style movement labels correspond to artistic movements such as "Impressionism" and "Expressionism". We use the WikiArt dataset [2] for evaluation, a popular artwork dataset with both genre and style movement annotations. The dataset comprises a total of 81,445 paintings: 57,025 in the training set, 12,210 in the validation set, and 12,210 in the test set, with three types of labels: 23 artists, 10 genres, and 27 style movements. All evaluation results are computed on the test set.

2.5.2 Training Data

Baselines reported on WikiArt are typically trained with the WikiArt training set. GOYA is trained with generated images by Stable Diffusion, which are described in the next paragraph. Additionally, the training dataset of Stable Diffusion LAION-5B [72] contains over five billion image–text pairs, which contain some paintings from the WikiArt test set. We examine other models trained on generated images, which are equally affected by this issue.

2.5.3 Image Generation Details

To generate images resembling human-made paintings, we relied on craft prompts $x = \{x^C, x^S\}$ as explained in Section 2.3.1. For simplicity, we selected titles of paintings as x^C and style movements as x^S, although alternative definitions of content and style descriptions could be used. In total, there are 43,610 content descriptions x^C, and 27 style descriptions x^S. For each x^C, we randomly selected five x^S to generate five prompts x. Then, each prompt generated five images with random seeds. In total, we obtained 218,050 prompts and 1,090,250 synthetic images. We split the generated images into 981,225 training and 109,025 validation images. We used Stable Diffusion v1.4[2] and generated images of size 512×512 through 50 PLMS [73] sampling steps.

Figure 2.3 depicts examples of diffusion generated images created by the specified prompts. We observed that the depicted scene is consistent with the content description in the prompts. Images in the same column have the same x^C but different x^S, exhibiting a high level of agreement in content while carrying significant differences in style. Likewise, images in the same row have the same x^S but different x^C, and paint different scenes or objects while maintaining a similar style. However, some content descriptions are religious, such as x^C in the third column, "our father who art in heaven". In such cases, achieving agreement on the semantic consistency between the generated images and the prompts may pose challenges.

2.5.4 GOYA Details

For the CLIP image and text encoders, we employ the pre-trained weights of CLIP-ViT-B/32 models[3]. The margin for computing contrastive losses is set to $\epsilon^C = \epsilon^S = 0.5$. In the indicator function for the content contrastive loss, the threshold ϵ^T is set to 0.25. We use the Adam optimizer [74] where base learning rate $= 0.0005$ and decay rate $= 0.9$. GOYA is trained on four A6000 GPUs with Distributed Data Parallel in PyTorch[4]. In each device, the batch size is

[2] https://github.com/CompVis/stable-diffusion

[3] https://github.com/openai/CLIP

[4] https://pytorch.org/

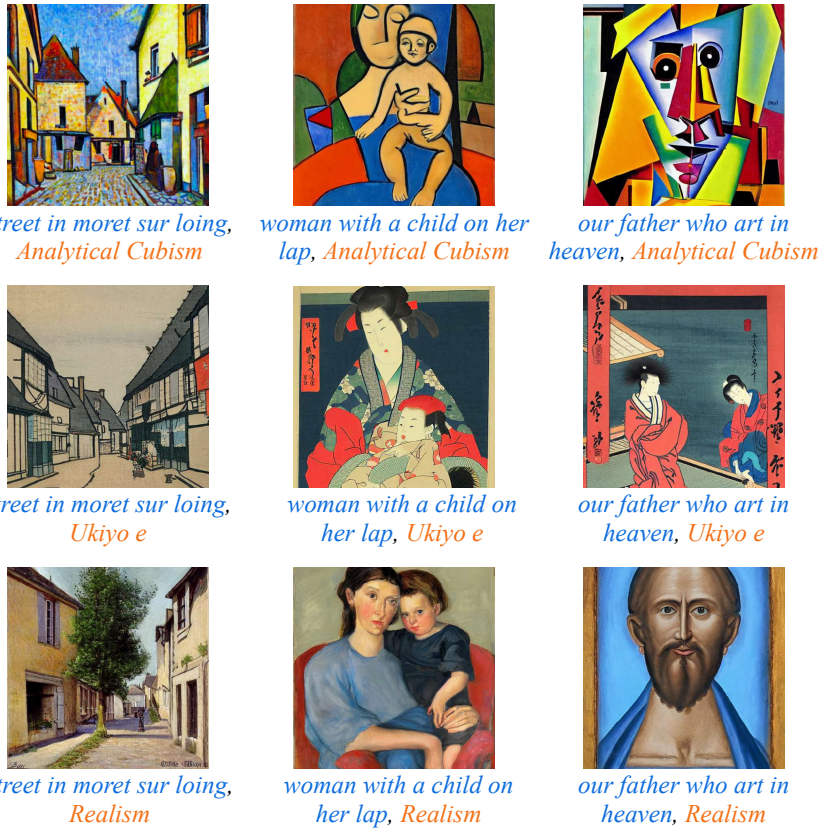

Figure 2.3: Examples of prompts and the corresponding generated diffusion images. The first part of the prompt (in blue) denotes the content description x^C, and the second part (in orange) is the style description x^S. Each column depicts the same content x^C while each row depicts one style x^S.

set as 512. Before being fed into CLIP, images are resized to 224×224 pixels. The architectural details of GOYA are shown in Table 2.1.

2.5.5 Disentanglement Evaluation

To measure content and style disentanglement quantitatively, we compute the distance correlation (DC) [75] between content and style embeddings, which is specially designed for content and style disentanglement evaluation. Let G^C and G^S denote matrices containing all content and style embeddings in the WikiArt test set, i.e., $G^C = (g_1^C \quad \cdots \quad g_N^C)$ and $G^S = (g_1^S \cdots \quad g_N^S)$.

Table 2.1: GOYA detailed architecture.

Components	Layer details
Content encoder $\mathcal{C}$	Linear layer $(512, 2048)$
	ReLU non-linearity
	Linear layer $(2048, 2048)$
Style encoder $\mathcal{S}$	Linear layer $(512, 512)$
	ReLU non-linearity
	Linear layer $(512, 512)$
	ReLU non-linearity
	Linear layer $(512, 2048)$
Projector h^C/h^S	Linear layer $(2048, 2048)$
	ReLU non-linearity
	Linear layer $(2048, 64)$
Style classifier $\mathcal{R}$	Linear layer $(2048, 27)$

For an arbitrary pair (i, j) of embeddings, the distances p_{ij}^C and q_{ij}^S can be computed by

$$p_{ij}^C = \|g_i^C - g_j^C\|, \quad p_{ij}^S = \|g_i^S - g_j^S\|, \tag{2.9}$$

where $\|\cdot\|$ gives the Euclidean distance. Let $\bar{p}_{i\cdot}^C$, $\bar{p}_{\cdot j}^C$, and $\bar{p}^C$ denote the means over j, i, and both i and j, respectively. With these means, the distances can be doubly centered by

$$q_{ij}^C = p_{ij}^C - \bar{p}_{i\cdot}^C - \bar{p}_{\cdot j}^C + \bar{p}^C, \tag{2.10}$$

and likewise for q_{ij}^S. DC between G^C and G^S is given by

$$\mathrm{DC}(G^C, G^S) = \frac{\mathrm{dCov}(G^C, G^S)}{\sqrt{\mathrm{dCov}(G^C, G^C)\mathrm{dCov}(G^S, G^S)}}, \tag{2.11}$$

where

$$\mathrm{dCov}(G^C, G^S) = \frac{1}{N}\sqrt{\sum_i \sum_j q_{ij}^C q_{ij}^S}. \tag{2.12}$$

$\mathrm{dCov}(G^C, G^C)$ and $\mathrm{dCov}(G^S, G^S)$ are defined likewise. DC can be computed for arbitrary matrices with N columns. DC is in $[0, 1]$, and a lower value means G^C and G^S are less correlated. We aim at DC being close to 0.

Baselines

To compute the lower bound DC on the WikiArt test dataset, we assigned the one-hot vector of the ground-truth genre and style movement labels as the content and style embeddings, representing the uppermost disentanglement when the labels are 100% correct. Besides the lower bound, we evaluated DC on ResNet50 [71], CLIP [26], and DINO [52]. For ResNet50, embeddings were extracted before the last fully connected layer. For CLIP, we used the embedding from the CLIP image encoder $\mathcal{E}_I$. For pre-trained DINO, following Splice [51], content and style embeddings were extracted from the deepest layer from the self-similarity of keys in the attention module and the $[\mathrm{CLS}]$ token, respectively.

Results

Results are reported in Table 2.2. With the lowest DC of 0.367, GOYA demonstrates the best disentanglement, surpassing the second-best model fine-tuned CLIP by a large margin. With

Table 2.2: Distance Correlation (DC) between content and style embeddings on the WikiArt test set. *Labels* indicate the results when using a one-hot vector embedding of the ground truth labels. ResNet50 and CLIP are fine-tuned on WikiArt, while DINO loads the pre-trained weights.

Model	Training Params.	Training Data	Emb. Size Content	Emb. Size Style	DC ↓
Labels	-	-	*27*	*27*	*0.269*
ResNet50 [71]	47M	WikiArt	2048	204	0.635
CLIP [26]	302M	WikiArt	512	512	0.460
DINO [52]	-	-	616,225	768	0.518
GOYA (Ours)	15M	Diffusion	2048	2048	**0.367**

only nearly $1/3$ training parameters of ResNet50 and $1/20$ of CLIP, GOYA outperforms embeddings directly trained on WikiArt's real paintings while consuming fewer resources. Also, GOYA achieves better disentanglement capability than DINO, with much more compact embeddings, e.g., $1/300$ content size embedding. However, there is still a notorious gap between GOYA and the lower bound based on labels, showing that there is room for improvement.

2.5.6 Similarity Retrieval

Next, we evaluate the visual retrieval performance of GOYA. Given a painting as a query, the five closest images are retrieved based on the cosine similarity of the embeddings in the content and style space, representing the most similar paintings in each space.

Results

Visual results are shown in Figure 2.6. Most of the paintings retrieved in the content space depict scenes similar to the query image. For instance, in the third query image, a woman with a headscarf is depicted bending over to scrub a pot, while all similar paintings in the content

space show a woman leaning to do manual labor such as washing, knitting, and chopping, independently of their visual style. It can be seen that, in most similar content paintings, various styles are depicted through different color compositions and tones. On the contrary, similar paintings in the style space tend to exhibit similar styles but different content. Similar images in the style space possess similar color compositions or brushstrokes, but depict distinct scenes compared to the query image. For example, the fourth query image, one of the paintings in the "*Rouen Cathedral*" series by Monet, exhibits different visual appearances on the same object under the light variance. It can be observed that the retrieved images in the style space also employ different light conditions to create a sense of space and display vivid color contrast. Furthermore, they also display similar color compositions and strokes but paint different scenes.

Figure 2.4 and 2.5 show results comparing against CLIP. In Figure 2.4 and 2.5, for each query image, the first two rows display the retrieved images from GOYA content and style spaces, and the last row shows images retrieved in the CLIP latent space. Results show that images in the CLIP latent space are similar in content and style, while in GOYA content space, there is consistency in depicting scenes but with different styles, and in GOYA style space, the visual appearance is similar, but the content is different.

2.5.7 Classification Evaluation

For evaluating the disentangled embeddings for art classification, following the protocol in [76], we trained two independent classifiers with a single linear layer on top of the content and style embeddings. We used 10 genres (genre labels include *abstract painting, cityscape, genre painting, illustration, landscape, nude painting, portrait, sketch and study, religious painting,* and *still life*) and 27 style movements (style movement labels include *Abstract Expressionism, Action painting, Analytical Cubism, Art Nouveau, Baroque, Color Field Painting, Contemporary Realism, Cubism, Early Renaissance, Expressionism, Fauvism, High Renaissance, Impressionism, Mannerism Late Renaissance, Minimalism, Naive Art Primitivism, New Realism, Northern Renaissance, Pointillism, Pop Art, Post Impressionism, Realism, Rococo, Romanticism, Sym-*

bolism, Synthetic Cubism and *Ukiyo-e*) in the WikiArt [2] dataset for classification evaluation.

Baselines

We compared GOYA against three types of baselines: pre-trained models, models trained on WikiArt dataset, and models trained on diffusion generated images. As pre-trained models, we used the Gram matrix [69, 77], ResNet50 [71], CLIP [26], and DINO [52]. For models trained on WikiArt, other than fine-tuning ResNet50 and CLIP, we also applied two popular contrastive learning methods: SimCLR [70] and SimSiam [76]. For models trained on generated images, ResNet50 and CLIP are fine-tuned with style movements in the prompts. When fine-tuning ResNet50 and CLIP, a linear classifier was added after the layer where embeddings are extracted, and we then trained the entire model on top of the pre-trained checkpoint. SimCLR and SimSiam were trained without any annotations.

Here we clarify the layer where the embeddings were extracted. Gram matrix embeddings are computed from the layer *conv5_1* of a pre-trained VGG19 [78]. For ResNet50 [71], CLIP [26], and DINO [52], the protocols for which layer to extract embeddings and for fine-tuning are consistent as in the disentanglement task.

Results

Table 2.3 shows the classification results. Compared with the pre-trained baselines listed in the first four rows, GOYA surpasses the Gram matrix, ResNet50, and DINO. However, it falls short of the pre-trained CLIP by less than 1% in both genre and style movement accuracy. Compared with models trained on WikiArt, although not comparable to fine-tuned ResNet50 and CLIP on classification, GOYA demonstrates superior disentanglement capabilities, as shown in Table 2.2. Moreover, GOYA exhibits enhanced classification performance when compared to contrastive learning models SimCLR and SimSiam.

When trained on diffusion generated images, GOYA achieves the best classification performance compared to other models with different embedding sizes. After fine-tuning on style movement in the prompts, ResNet50 shows a 3% increase on the style accuracy, indicating the

Table 2.3: Genre and style movement accuracy on the WikiArt [2] dataset for different models.

Model	Training Data	Label	Num. Train	Emb. Size Content	Emb. Size Style	Accuracy Genre	Accuracy Style
Pre-trained							
Gram Matrix [69, 77]	-	-	-	4096	4096	61.81	40.79
ResNet50 [71]	-	-	-	2048	2048	67.85	43.15
CLIP [26]	-	-	-	512	512	**71.56**	**51.23**
DINO [52]	-	-	-	616,225	768	51.13	38.81
Trained on WikiArt							
ResNet50 [71] (Genre)	WikiArt	Genre	57,025	2048	2048	79.13	43.17
ResNet50 [71] (Style)	WikiArt	Style	57,025	2048	2048	67.22	**64.44**
CLIP [26] (Genre)	WikiArt	Genre	57,025	512	512	**80.43**	34.98
CLIP [26] (Style)	WikiArt	Style	57,025	512	512	56.28	63.02
SimCLR [70]	WikiArt	-	57,025	2048	2048	65.82	45.15
SimSiam [76]	WikiArt	-	57,025	2048	2048	51.65	31.24
Trained on Diffusion generated							
ResNet50 [71] (Movement)	Diffusion	Movement	981,225	2048	2048	61.78	45.79
CLIP [26] (Movement)	Diffusion	Movement	981,225	512	512	52.65	43.58
SimCLR [70]	Diffusion	-	981,225	2048	2048	33.82	20.88
GOYA (Ours)	Diffusion	-	981,225	2048	2048	**69.70**	**50.90**

potential for analysis via synthetically generated images. However, CLIP decreases in both genre and style accuracy after fine-tuning on generated images. SimCLR experiences a dramatic decrement when trained on generated images compared to WikiArt. As SimCLR focuses more on learning the intricacies of the image itself rather than the relation of images, it learns the distribution of generated images, leading to poor performance on WikiArt. While training on the same dataset, GOYA maintains better capability on classification tasks while achieving high disentanglement.

To thoroughly examine the classification results, we provide confusion matrix analyses for both genre and style movement classification evaluations. Figure 2.7 shows the confusion matrix of genre classification evaluation on GOYA's content space. The number in each cell rep-

resents the proportion of images classified as the predicted label to the total images with the true label. The darker the color, the more images are classified as the predicted label. We can observe that images from several genres are misclassified as *genre painting*, as such paintings usually depict a wide range of activities in daily life, thus overlapping semantically with images from other genres, such as *illustration* and *nude painting*. In addition, due to the high similarity of depicted scenes, there is a 28% misclassification rate of images from *cityscape* as *landscape*.

The confusion matrix of style movement classification is shown in Figure 2.8. However, the boundary of some movements is not very clear, as some movements are sub-movements that represent different phases within one major movement, e.g., *Synthetic Cubism* in *Cubism* and *Post Impressionism* in *Impressionism*. Generative models may produce images likely to belong to the major movement even when the prompt is about sub-movements, leading GOYA to learn from inaccurate information. Thus, images from sub-movements are prone to be predicted as the according major movement. For example, 82% of the images in *Synthetic Cubism* and 90% of the images in *Analytical Cubism* are classified as *Cubism*. Similarly, about $1/3$ of the images in *Contemporary Realism* and *New Realism* are predicted incorrectly as *Realism*.

2.5.8 Ablation Study

We conducted an ablation study on the WikiArt test set to assess the effectiveness of the losses and the network structure in GOYA.

Losses

We compare the losses used in GOYA against two other popular contrastive losses, Triplet loss [79] and NTXent loss [80], both of which have shown their superiority in many contrastive learning methods. We also investigated the application of a style classification loss in conjunction with the above-mentioned contrastive losses. The criteria of selecting positive and negative pairs remain consistent across all of these loss functions.

The results in terms of accuracy (as the product of genre and style movement accuracies) and disentanglement (as DC) are depicted in Figure 2.9. The NTXent loss achieves the highest

accuracy but with the cost of undercutting the disentanglement ability. In contrast, Triplet loss exhibits almost the best disentanglement performance but lags behind in terms of classification performance. Compared to these two losses, only the contrastive loss in GOYA manages to maintain a balance between disentanglement and classification performance. Moreover, after occupying the classification loss, GOYA has a boost in classification accuracy without sacrificing disentanglement, achieving the best performance compared to the other loss settings.

Embedding Size

We explore the effect of the embedding size on a single-layer content and style encoders, ranging from 256 to 2048. Figure 2.10 illustrates that both genre and style accuracy improve by up to 6% as the embedding size increases, but conversely, the DC deteriorates, from 0.750 to 0.814, indicating a trade-off between classification and disentanglement. Moreover, the classification performance of genre and style movement surpasses the pre-trained CLIP (shown in Table 2.3) when the embedding size exceeds 512, suggesting that larger embedding sizes possess a stronger ability to distill knowledge from the pre-trained model. Inspired by this finding, we set the embedding size to 2048.

2.6 Discussion

2.6.1 Image Generation

- **Prompt design**: In this study, we used a combination of content and style descriptions as prompts, where the content description comprises the title of paintings, and the style description employs the style labels of the WikiArt dataset. Alternatively, more specialized prompt designs could be implemented to attain even finer control over the generated images. For example, captions from vision-language datasets could be employed as content descriptions, while detailed style descriptions could be extracted from external knowledge such as Wikipedia.

- **Data replication**: As demonstrated in previous research [81, 82], Stable Diffusion might produce forgeries, generating images that closely resemble the training data. However, the extent of these replicated images within our training data remains uncertain, and their potential impact on model training has yet to be thoroughly explored.

2.6.2 Model Training

- **Encoder structure**: For the content and style encoders, we employ small networks consisting of only two and three layers, respectively. We found that a higher-dimensional hidden layer (2048) and fewer layers (3) are effective for learning content embeddings, while a lower-dimensional hidden layer (512) and more layers (2) yield better style embeddings. We hypothesize that content embedding, which reflects semantic information, benefits from a large number of neutrons, while style embedding, containing low-level features, is more efficiently represented with lower dimensions.

- **Partition of synthetic images**: We performed style movement classification on a training dataset comprising both synthetic and real data. Results presented in fig. 2.11 indicate that, as the number of synthetic images increases during training, the accuracy decreases. We attribute this phenomenon to the domain gap between synthetic and real images. In addition, we suggest that contrastive learning may help alleviate the impact of this domain gap.

2.6.3 Limitation on the WikiArt Dataset

While the WikiArt dataset serves as our evaluation dataset, it comes with limitations related to annotations and diversity. Firstly, the annotated genre and style movement label may not entirely align with the content and style definitions described in this paper. Secondly, the majority of the paintings in WikiArt belong to Western art, especially European and American art, thus lacking representation from a diverse spectrum of art paintings. Future work could focus on obtaining

more precise annotations for content and style in paintings, as well as including art paintings from various regions, such as Asian, Oceanian, and African art, thereby enriching the diversity of the dataset.

2.6.4 Applications

- **Art applications**: Our work can potentially be extended into various practical scenarios. For instance, it could be integrated into an art retrieval system, enabling users to find paintings based on text descriptions or a given artwork. Additionally, it could be employed in a painting recommendation system, offering personalized suggestions to users according to their preferred paintings. These applications have the potential to enhance user experience and engagement, thus contributing to the improvement of art production and consumption.

- **Digital humanities**: While our work mainly focuses on the analysis of fine art, there is potential for our work to be applied in other areas within digital humanities, such as graphic design and historical document analysis.

- **Beyond art**: Apart from the art domain, audio disentanglement could be a potential area to expand [83–85].

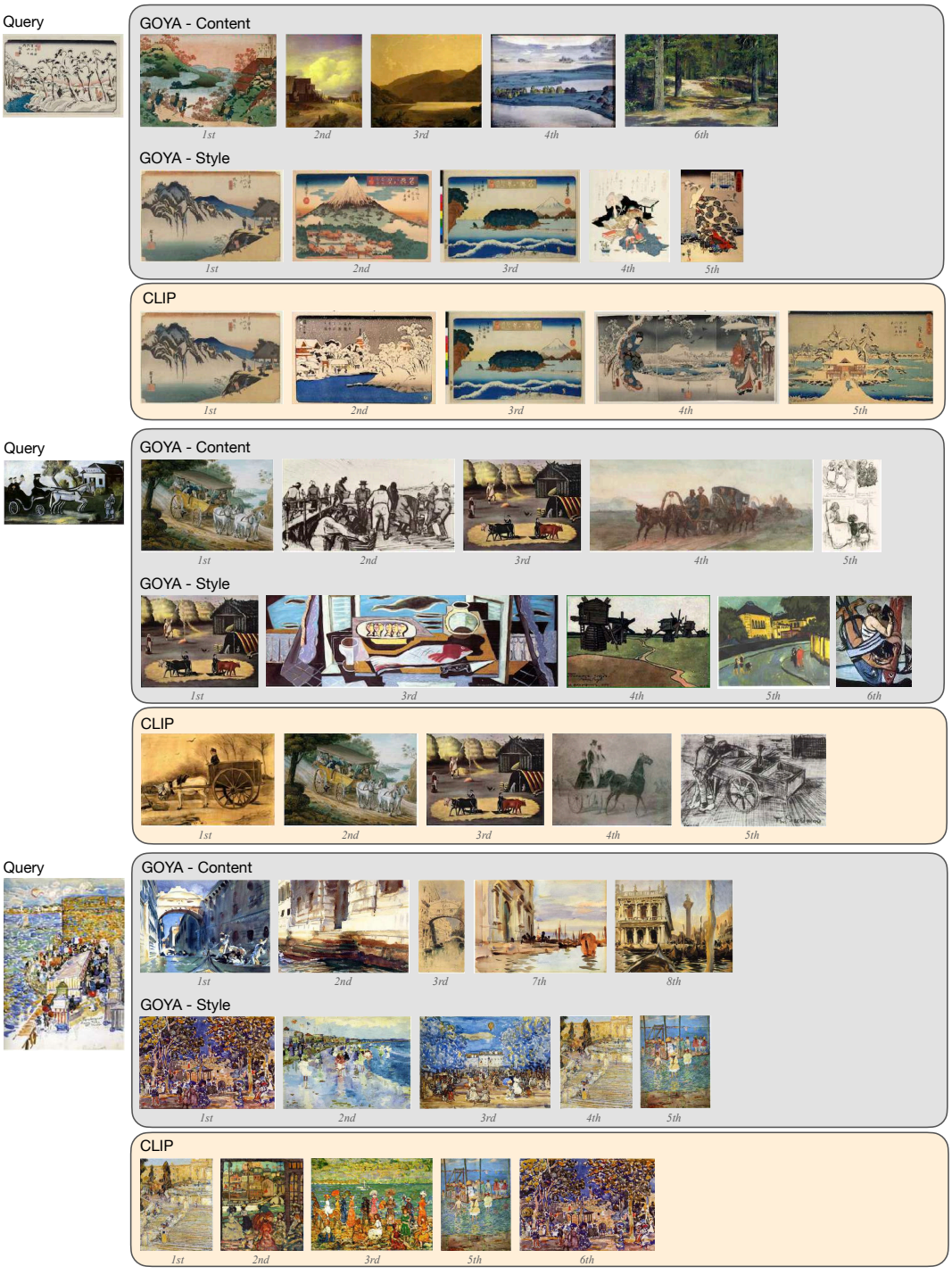

Figure 2.4: Retrieval results in GOYA content and style spaces and CLIP latent space based on cosine similarity. In each row, the similarity decreases from left to right. Copyrighted images are skipped.

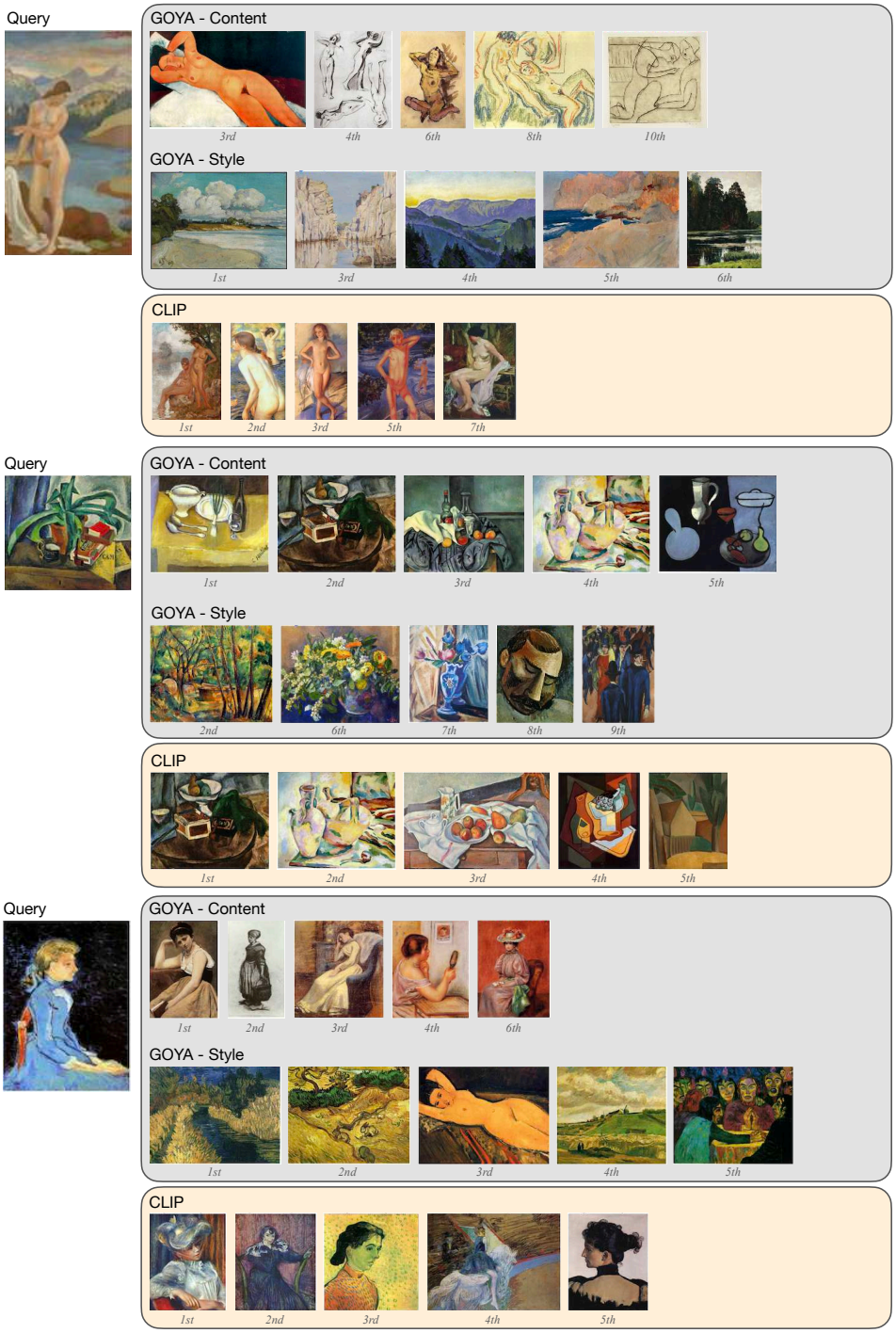

Figure 2.5: More retrieval results in GOYA content and style spaces and CLIP latent space based on cosine similarity. In each row, the similarity decreases from left to right. Copyrighted images are skipped.

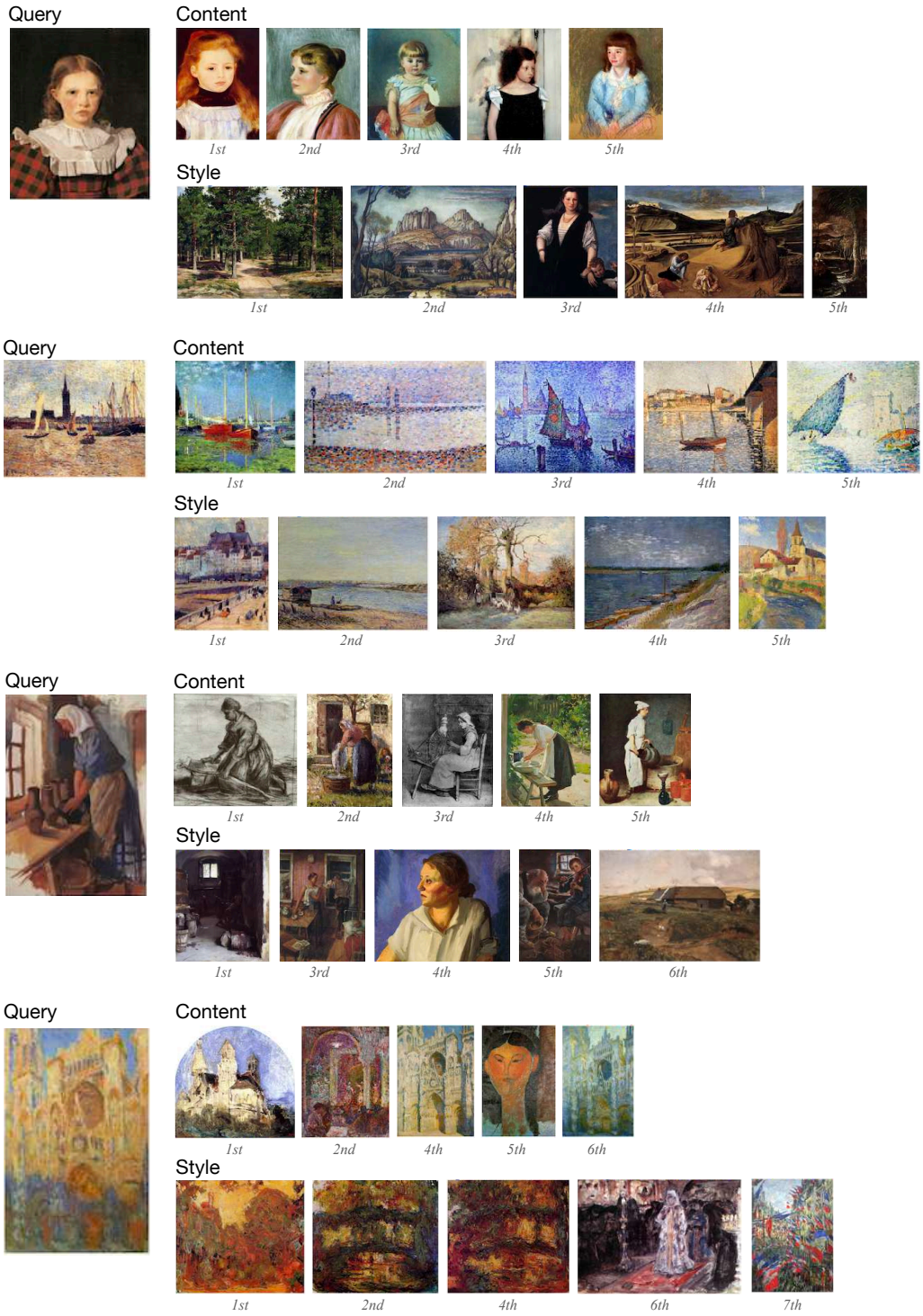

Figure 2.6: Similarity retrieval in the content and style spaces using GOYA on the WikiArt test set. The similarity decreases from left to right. Copyrighted images are skipped.

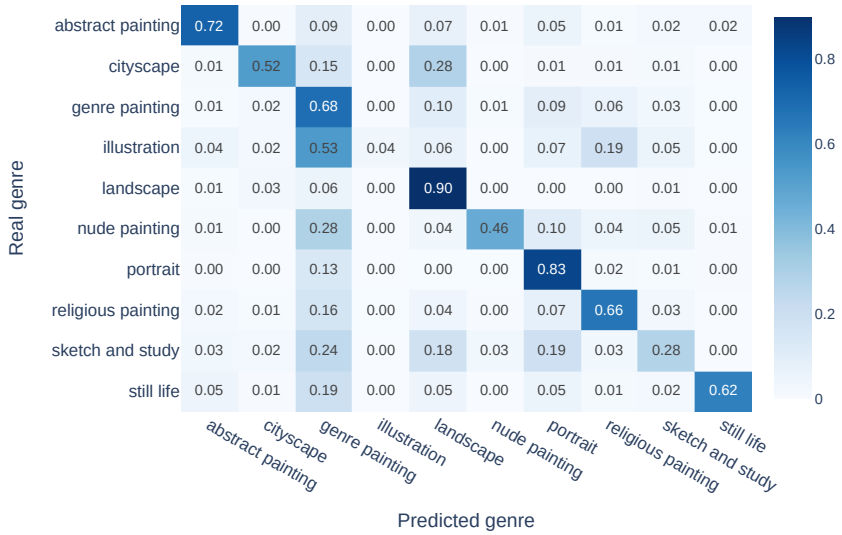

Figure 2.7: Confusion matrix for genre classification evaluation in the content space using GOYA.

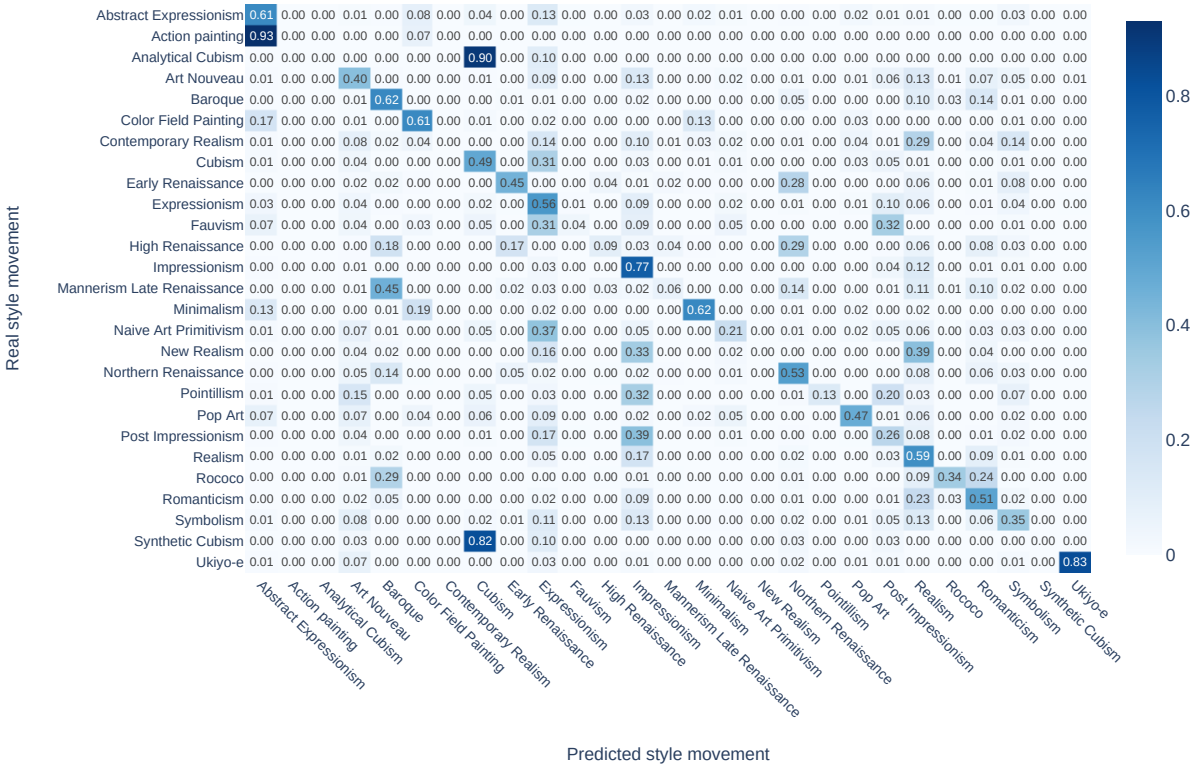

Figure 2.8: Confusion matrix for style movement classification evaluation in the style space using GOYA.

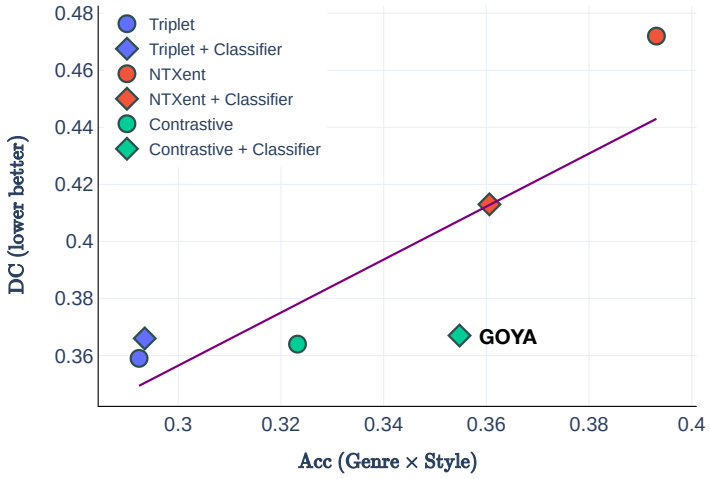

Figure 2.9: Loss comparison. The x-axis shows the product of genre and style accuracies (the higher the better) while the y-axis presents the disentanglement, DC (the lower the better). The purple line shows the trendline as $y = 0.0776 + 0.9295x$. In general, better accuracy is obtained at expense of a worse disentanglement. Only GOYA (Contrastive + Classifier loss) improves accuracy without damaging DC.

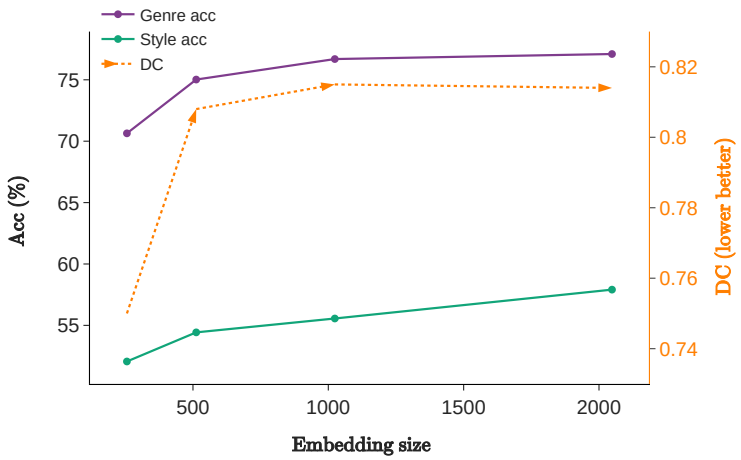

Figure 2.10: Disentanglement and classification evaluation with different embedding sizes when only one single layer is set in the content and style encoder. A larger embedding size benefits the genre and style movement accuracy but leads to worse disentanglement.

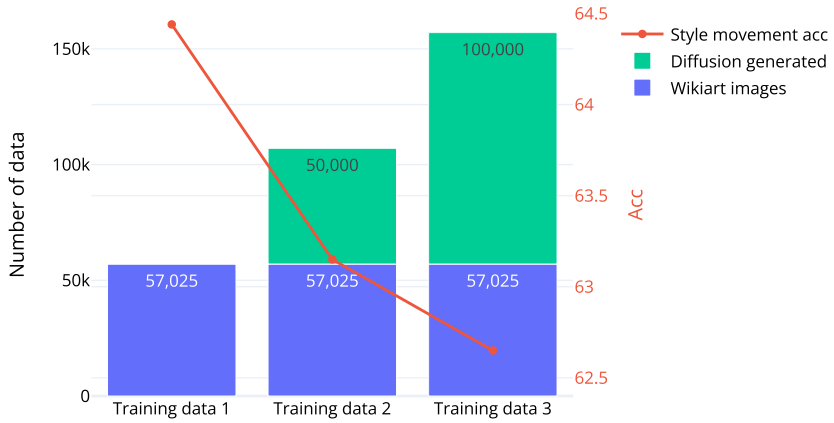

Figure 2.11: Style classification on ResNet50 when the training set contains both synthetic and real data. As the partition of synthetic images increases, the style movement accuracy drops.

Chapter 3

Revealing Gender Bias from Prompt to Image in Stable Diffusion

3.1 Overview

Text-to-image generation has shown a superior capability for generating high-fidelity images. Given natural language inputs, known as prompts, cutting-edge models such as Stable Diffusion [3] and DALL-E 2 [4] produce high-quality images that align closely with the given prompts. However, their widespread accessibility and diverse applications across various domains have raised ethical concerns, such as the social impact of data [86–88], bias [10, 89, 90], privacy [59, 82], or intellectual property issues [81, 91]. Evaluating these problems remains a relatively underexplored challenge. In this work, we focus on developing an evaluation protocol for gender bias in Stable Diffusion models.

It has been widely shown that certain adjectives [89] or professions [89] can lead to the generation of stereotypical demographic attributes in faces. However, disparities according to gender are also shown in regions beyond the faces, which are intended to fill the images [10]. Figure 3.1 shows triplets of generated images from prompts that differ only in the gender indicators (gender indicators refer to words that indicate the gender of a person). We observe

Figure 3.1: We use free-form triplet prompts to analyze the influence of gender indicators on the overall image generation process. We show that (1) gender indicators influence the generation of objects (**left**) and their layouts (**right**), and (2) the use of gender *neutral* words tends to produce images more similar to those prompted by *masculine* indicators rather than *feminine* ones.

that while the representation of the faces changes accordingly, unexpected variations also occur in other parts of the images, even when not explicitly mentioned in the prompt. For example, differences can be seen in the object depicted (e.g., different musical instruments on the upper-left image) and the layout of the image (e.g., on the right images). This suggests that gender bias extends beyond face representations and influences the broader context of the entire image. Most previous works [10, 87, 89, 90, 92–95] report demographic bias focused on generated faces in text-to-image generation [10, 89, 92, 96], often neglecting to examine the generation process and how bias perpetuates from prompt to image.

In this paper, we investigate the internal components of Stable Diffusion to uncover the origins of gender bias and how it pertains [97, 98]. We suggest that these disparities arise from the interplay of representational disparities and prompt-image dependencies during image generation: the process involves transitioning from prompt space to image space, potentially treating genders differently and resulting in representational disparities. To analyze differences regard-

ing genders, we set triplet prompts that differ only in gender indicator, and quantify representational disparities (Section 3.4) and prompt-image dependencies (Section 3.6). Our automatic evaluation protocol allows us to formulate and answer the following research questions (RQ):

RQ1 Do images generated from neutral prompts exhibit greater similarity to those generated from masculine prompts than to images generated from feminine prompts and, if so, why?

RQ2 Do object occurrences in images significantly vary based on the gender specified in the prompt? If there are differences, do these object occurrences from neutral prompts exhibit greater similarity to those from masculine or feminine prompts?

RQ3 Does the gender in the input prompt influence the prompt-image dependencies in Stable Diffusion, and if so, which prompt-image dependencies are more predisposed to be affected?

We conduct experiments on three versions of Stable Diffusion models. The template-free natural language prompts are derived from four caption datasets and a text set generated by ChatGPT [12]. Despite differing only in the gender indicator, the triplets exhibit a consistent trend across all Stable Diffusion models. Our key findings include the following:

- The images generated from neutral prompts are consistently more similar to those from masculine prompts than feminine prompts.
- Across *all internal stages of the generation process*, representation from neutral prompts also exhibits greater similarity to those from masculine than from feminine ones.
- Object co-occurrence in images generated from neutral prompts aligns more closely with masculine prompts than with feminine prompts.
- Objects explicitly mentioned in the prompts do not exhibit differences regarding specific gender.
- Objects not explicitly mentioned in the prompts have different possibilities to be generated regarding different genders.

These findings demonstrate that gender bias perpetuates throughout the generating process and manifests across entire images, including areas beyond generated faces. To address this issue, we provide recommendations for both model developers and users to mitigate bias during image generation. Compared to our conference version [97], this work includes the following improvements[1]:

- An extended literature review on gender bias evaluation methods in text-to-image generation.
- Additional details on triplet prompt generation (section 3.3.1), image space (section 3.4.1), and word attention (section 3.6.1).
- Expanded experimental results and further discussions in Sections 3.4–3.6.
- Deeper analysis of the prompt-image dependency, including dependency group presence in images (section 3.7.2), amount of objects (section 3.7.2).

3.2 Related Work

3.2.1 Text-to-Image Models

There are three main types of text-to-image generation models: GAN [57, 99, 100], autoregressive [4, 53, 101–103], and diffusion [3, 65, 104]. Within diffusion models, Stable Diffusion [3] has emerged as the preferred testbed due to its high-quality generations and open-source nature. As diffusion models rely on cross-attention to connect text and image modalities, it enables the examination of the image generation process at the word level [105]. The cross-attention module assists in tasks such as editing [105–109] and segmentation [110–112]. By leveraging this property, we can investigate the relationship between gender and prompt-guided generations.

[1]This chapter is based on the conference paper [97].

3.2.2 Social Bias

Text-to-image generation models often reproduce demographic stereotypes tied to gender and race across various factors, including but not limited to occupations [10, 89, 92–94, 96, 113], adjectives [89, 95, 114], objects [115], outfits [116], and nationalities [10, 117]. Analysis of prompt templates like "`a photo of the face of [OCCUPATION]`" reveals that certain occupations, such as *software developers*, are predominantly represented as white men, while *housekeepers* tend to be associated with women of color. Additionally, Wolfe et al. [11] showed that models are more inclined to generate sexualized images in response to prompts containing "`a [AGE] year-old girl`". Moreover, Zhang et al. [118] argued that unfairness extends to images depicting underrepresented attributes like *wearing glasses*, highlighting the pervasive nature of biases in the generation process. In addition to biases concerning humans, previous studies have explored geographical-level differences in objects [119] and the correctness of cultural context [120, 121].

3.2.3 Bias Evaluation

A fundamental aspect in the study of bias is the evaluation protocol. As summarized in Table 3.1, we compare differences between our method and several previous gender bias evaluation methods in text-to-image generation [8–10, 87, 89, 92, 93, 95, 115, 116, 122–131]. Most of these approaches rely on prompts that fill attributes (e.g., profession) with a template, leading to constrained scenarios and limited additional details in the prompts. Moreover, these methods evaluate bias on the proxy presentation of the generated images, but do not examine presentations in the generation process. Additionally, these methods mainly focus on people's attributes, such as the gender of faces, thereby overlooking biases in the generated visual elements as well as the entire image context. Except for the method that exclusively on gender bias evaluation, there are traditional evaluation criteria for text-to-image models measuring image fidelity and text-image alignment with automated metrics [132–135] or human evaluation [136].

Overall, there is an absence of automated methods for nuanced bias evaluation that conveys

bias at the different stages of the generation process. Using free-form prompts, our work proposes a method to uncover prompt-image dependencies, disclosing how objects are generated differently according to gender indicators in the prompt.

3.3 Preliminaries

3.3.1 Triplet Prompt Generation

Let $\mathcal{P}_{\mathrm{n}}$ be a set of *neutral* prompts, which do not specify the gender of the person. As shown in Figure 3.1, from these neutral prompts, we generate two counterpart prompt sets, $\mathcal{P}_{\mathrm{f}}$ and $\mathcal{P}_{\mathrm{m}}$, as *feminine* and *masculine* prompt sets, respectively. The only difference among these three prompt sets is the gender indicator, while all other words remain unchanged. Our bias evaluation is based on analyzing distinctions between pairs of generated images from the triplet $\{\mathcal{P}_{\mathrm{n}}, \mathcal{P}_{\mathrm{f}}, \mathcal{P}_{\mathrm{m}}\}$.

We generate neutral prompts from natural language sentences, consisting of captions from four vision-language datasets (GCC validation set [140], COCO [137], TextCaps [141], and Flickr30k [139]), as well as a profession prompt set generated by ChatGPT 3.5 [12] (accessed on 7 November 2023). From the vision-language datasets, we generate neutral prompts by choosing *neutral captions*. To ensure the neutral prompts do not contain other words that might potentially define the gender of generated people, we set two criteria for the neutral captions: (1) they contain the word *person* or *people*, and (2) they do not include other words (listed in Table 3.2) that indicate humans (e.g., "`The person and a boy are playing badminton`" is not a neutral caption). To generate feminine and masculine prompts, we swap *person*/*people* in the neutral captions with the gender indicators *woman*/*women* and *man*/*men*, respectively. For the profession prompt set, we generate neutral prompts with ChatGPT based on professions, such as `ecologist` or `doctor`, across 16 topics. For example, `an ecologist studies the ecosystem in a lush green forest.` To create feminine and masculine prompts, we prepend *female/male* before the profession (e.g., `an` *`female`* `ecologist studies the`

`ecosystem in a lush green forest)`. Examples of triplet prompts and the corresponding generated images for each dataset are shown in Figure 3.8.

3.3.2 Image Generation

Given prompt p as input, Stable Diffusion transforms it into a text embedding $\mathbf{t}$ in the *prompt* space using the text encoder. This text embedding is fed into the cross-attention module in UNet [142], which performs the denoising operations from an initial noise $\mathbf{z}_T$ in the latent space. After T denoising steps, the embedding $\mathbf{z}_0$ in the *denoising* space is obtained. Finally, image x in the *image* space is generated from $\mathbf{z}_0$ by the image decoder. In this work, we evaluate Stable Diffusion models: v1.4[2], v2.0-base[3], and v2.1-base[4] (denoted as SD v1.4, SD v2.0, and SD v2.1, respectively). The three versions share the same model structure as introduced, but they differ in their text encoders. SD v1.4 uses CLIP ViT-L/14, while SD v2.0 and SD v2.1 use a larger and more transparent encoder, OpenCLIP ViT-H.

Table 3.3 reports the details of image generation for each dataset. The seed is the same within each triplet, ensuring the same initial noise $\mathbf{z}_T$. To address data scarcity in GCC and Profession sentences, we produce five images per prompt with five different seeds. In the following, when mentioning a dataset, we are referring to the generated images whose prompts originate from the corresponding dataset.

3.3.3 Gender Bias Definition

The interpretation of gender bias varies across literature, resulting in different work attributing different meanings to the term. In this paper, we define gender bias as follows:

- Within the triplet, images generated from *neutral* prompts consistently display greater similarity to those from either *feminine* or *masculine* prompts.

[2]`https://github.com/CompVis/stable-diffusion`
[3]`https://huggingface.co/stabilityai/stable-diffusion-2-base`
[4]`https://huggingface.co/stabilityai/stable-diffusion-2-1-base`

- Specific objects tend to appear more frequently in the generated images associated with a specific gender.

Whereas objects are not equally distributed in the real world or across cultures, and recognizing that not all disparities regarding genders are inherently problematic (i.e., the association of *dress* with *women* may not be an issue, whereas *kitchen* might), we argue that it is essential to have a methodology for recognizing and quantifying these differences. Our proposed evaluation protocol is not envisaged to identify objects that perpetuate discrimination and gender stereotypes, but to *highlight significant gender disparities*, regardless of whether they are deemed problematic.

3.4 Gender Disparities in Neutral Prompts

RQ1 *Do images generated from neutral prompts exhibit greater similarity to those generated from masculine prompts than to images generated from feminine prompts and, if so, why?*

In this section, we address the above research question through the use of representational disparities.

3.4.1 Representational Disparities

We use representational disparities to analyze how images generated by different gender indicators compare with respect to neutral prompts. For a given triplet, the analysis consists on comparing the similarity between *neutral* embeddings and *feminine* and *masculine* embeddings. To measure the extent of gender disparities in the generative process, as shown in Figure 3.2, we examine the representational disparities throughout the entire generation, tracking embeddings from the prompt space to the denoising space and the image space, offering insights into when bias is introduced.

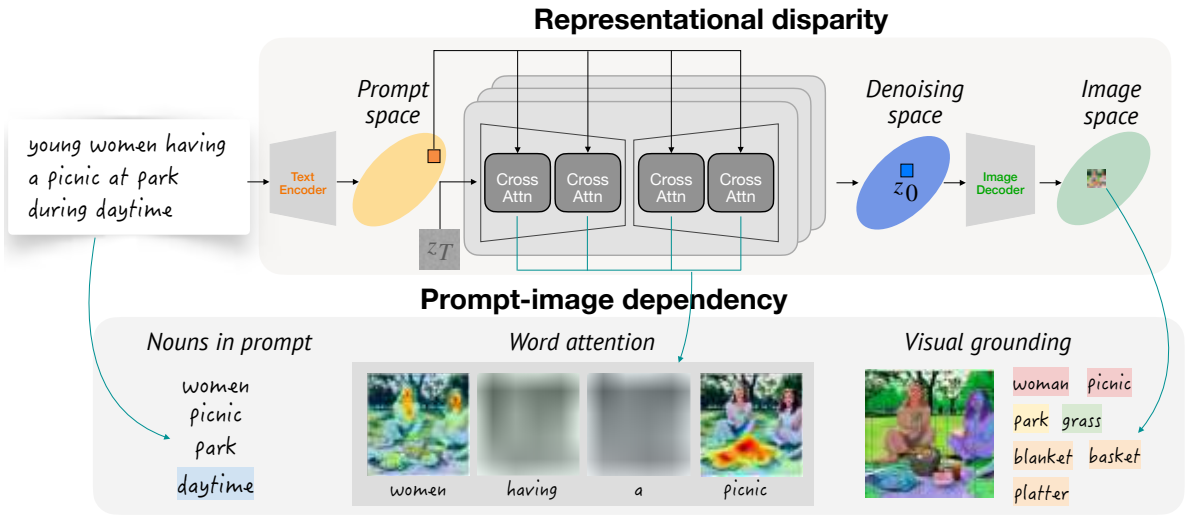

Figure 3.2: Overview of representational disparities and prompt-image dependency.

Prompt Space

The prompt space is defined as the space in which all text embeddings lie. Different points in this space provide different semantics to the following image generation process. To measure the disparity between a pair prompt set $\mathcal{P}$ and $\mathcal{P}'$ in the triplet, we compute cosine similarity as

$$s_{\mathrm{P}}(\mathcal{P}, \mathcal{P}') = \frac{1}{|\mathcal{P}|} \sum_{p_i, p_i'} \cos(\mathbf{t}, \mathbf{t}'), \tag{3.1}$$

where $|\cdot|$ is the number of elements in the given set, $\cos(\cdot, \cdot)$ gives cosine similarity, the summation is computed over all prompts p_i from $\mathcal{P}$ and p_i' from $\mathcal{P}'$ (subscript i is the index of the prompt to clarify p_i and p_i' are corresponding prompts, derived from the same one), and text embeddings $\mathbf{t}$ and $\mathbf{t}'$ correspond to prompts p_i and p_i', respectively.

Denoising Space

The embedding $\mathbf{z}_0$ after the last denoising process lies in the denoising space. Similarly to the prompt space, we compute cosine similarity as

$$s_{\mathrm{D}}(\mathcal{P}, \mathcal{P}') = \frac{1}{|\mathcal{P}|} \sum_{p_i, p_i'} \cos(\mathbf{z}_0, \mathbf{z}_0') \tag{3.2}$$

where $\mathbf{z}_0$ and $\mathbf{z}'_0$ are derived from p_i and p'_i, respectively.

Image Space

As bias often involves more in the semantics rather than pixel values, we adopt a spectrum of metrics computed from the generated images. To measure image structural differences, we use the average of SSIM scores over all pixels as one of our disparity metrics *SSIM*. Additionally, the ratio of the number of pixels in the contours with higher SSIM scores is used as another disparity metric *Diff. Pix.* To quantify differences in higher-level semantics, we apply latent vectors of pre-trained neural networks, adopting the last fully connected layer of ResNet-50 [71], the image encoder from CLIP ViT-B/32[5] [26], and the last layer of DINO-s16 [52] following [143], referred to as *ResNet*, *CLIP*, and *DINO*, respectively. For all metrics, we compute the cosine similarity between the latent vectors from image pairs as in Equations (3.1) and (3.2). Additionally, we adopt *split-product* [81] using DINO-b8 [52] following the default configuration, computing the maximum cosine similarity among corresponding patches between image pairs.

3.4.2 Results Analysis

By analyzing the representational disparities on (*neutral*, *feminine*), and (*neutral*, *masculine*) pairs, we can provide some answers for **RQ1**.

Results are shown in Table 3.8. In the image space, regardless of whether considering the entire image holistically (*SSIM*, *Diff. Pix*, *ResNet*, *CLIP*, and *DINO*), or the highest similarity on corresponding patches (*split-product*), images generated from *neutral* prompts consistently demonstrate greater similarity to those from *masculine* prompts. This trend is consistently observed in all datasets and all models.

Tracing back to the prompt space and denoising space to explore where and when gender bias emerges in the generated images, embeddings from *neutral* prompts are closer to the embeddings from *masculine* prompts, both in the prompt space and the denoising space. Although

[5] https://github.com/openai/CLIP

Stable Diffusion models apply different text encoders (OpenCLIP-ViT/H for SD v2.0 and SD v2.1, while CLIP ViT-L/14 for SD v1.4), the same trend is observed across all three models and all datasets. This indicates that gender bias originates from text embedding and perpetuates through the generation process, leading to the disparities observed in the generated images.

3.5 Influence of Gender on Objects

RQ2 *Do object occurrences in images significantly vary based on the gender specified in the prompt? If there are differences, do these object occurrences from neutral prompts exhibit greater similarity to those from masculine or feminine prompts?*

The representational disparities reflect the holistic similarity between gender groups, but they do not convey fine-grained differences, i.e., why a certain object appears in the generated image given a gender-specific prompt. In this section, we address **RQ2** by investigating the relationship between gender and the objects in the generated images. To do so, we extract objects with a visual grounding model and study their co-occurrence with each gender.

3.5.1 Detecting Generated Objects

To detect objects in the generated images we use the assembled model Grounded-SAM [144]. Given a generated image, RAM (14M) [145] predicts plausible objects, which are used by Grounded DINO-T [146] to propose bounding boxes around the candidate objects. Then, ViT-H Segment Anything Model (SAM) [147] extracts object regions m_o within the bounding box of the object o. For each image, a set of object names and a set of regions are obtained.

3.5.2 Evaluation Metrics

Our evaluation protocol involves measuring the differences in object co-occurrences for different genders. Let $\text{cnt}(o, p)$ denote the number of occurrences of the object o in the image generated from the prompt p in the prompt set $\mathcal{P}$. The total number of co-occurrence $C(o, \mathcal{P})$

is given by

$$C(o, \mathcal{P}) = \sum_{p \in \mathcal{P}} \text{cnt}(o, p) \tag{3.3}$$

With the above definition and a set of triplet prompts, we use the following three methods to evaluate the influence of gender in the generated objects.

(1) Statistical Tests We use the chi-square test to check whether there are statistical differences in the object co-occurrence among two or three image sets. This test is applicable to the triplet and any pairs in the triplet. If the resulting p-value is below 0.05, we interpret significant differences in the object distribution in the pair or triplet.

(2) Co-occurrence Similarity We compute the similarity of the co-occurrences of detected objects between two image sets. Formally, let the vector $\mathbf{v}_p$ denote the object occurrences in the image generated from prompt p, and each element in $\mathbf{v}_p$ is the occurrence $\text{cnt}(o, p)$ for the object o in the image. Similarly to Equations (3.1) and (3.2), we compute cosine similarity on object co-occurrences as

$$s_{\text{O}}(\mathcal{P}, \mathcal{P}') = \frac{1}{|\mathcal{P}|} \sum_{p_i, p'_i} \cos(\mathbf{v}_i, \mathbf{v}'_i), \tag{3.4}$$

where prompt sets $\mathcal{P}$ and $\mathcal{P}'$ are in the triplet. $\mathbf{v}_i$ and $\mathbf{v}'_i$ are derived from prompt p_i in $\mathcal{P}$ and p'_i in $\mathcal{P}'$, respectively. A higher co-occurrence similarity means that objects are detected with the same-level frequency in two image sets, whereas a low similarity means that objects are detected at different rates.

(3) Bias Score Following [148], we compute the bias score $\text{BS}(o)$ for a certain object o as

$$\text{BS}(o) = \frac{C(o, \mathcal{P}_{\text{m}})}{C(o, \mathcal{P}_{\text{m}}) + \frac{|\mathcal{P}_{\text{m}}|}{|\mathcal{P}_{\text{f}}|} C(o, \mathcal{P}_{\text{f}})}. \tag{3.5}$$

$\text{BS}(o)$ ranges from 0 to 1, with 1 meaning the object is skewed towards *masculine* prompts and 0 towards *feminine* prompts. If $\text{BS}(o) = 0.5$, object o does not favor any gender.

3.5.3 Results Analysis

All the p-values from chi-square tests among the triplets and pairs are below 10^{-5}, implying significant differences in the object distributions of each gender across all datasets and models. This shows that according to gender, not only the person in the image may change, but also the objects generated in the image are statistically different.

To investigate whether the object co-occurrences of neutral images exhibit larger similarity to a certain gender image set, we compute co-occurrence similarity on pairs (*neutral, feminine*) and (*neutral, masculine*). Results in Table 3.4 indicate that object co-occurrences in *neutral* consistently exhibit greater similarity to those in *masculine* prompts than in *feminine* prompts across all datasets and models, corroborating the observations in Section 3.4. This, again, indicates that prompts that use gender neutral words tend to generated objects that are more commonly generated for masculine prompts than for feminine prompts.

Subsequently, we examine specific examples by computing the bias score based on co-occurrence for each object in the generated images. We filter objects if the maximum co-occurrence is less than 10 in GCC, 20 in COCO, TextCaps, and Flickr30k, and 5 in Profession. Results are shown in Figure 3.3. We can observe that results exhibit a consistent trend across different datasets and models. Take SD v2.0 as an example, notably, clothing and accessory exhibit a high bias: for example, `suspender` (1 in GCC, Flickr30k, and Profession), `suit` (GCC, TextCaps, and Flickr30k (0.98), COCO (0.96)), and `bow tie` (GCC (0.96), COCO and TextCaps (0.98), Flickr30k (1)) lean towards *masculine*, while `bikini top` (GCC (0.05), COCO (0.01), Flickr30k (0.02)), `legging` (GCC (0.08), Flickr30k (0.02), COCO (0.01)), and `earring` (GCC (0.03), COCO (0.02), Profession (0)) lean towards *feminine*. This is not surprising, considering that clothing elements are traditionally gendered. Other than clothing, we find a strong association between `family` (0.11) and `child` (0.31) with *feminine* prompts, potentially associating *feminine* with caregiver, while *masculine* prompts exhibit greater alignment with words related to sports such as `baseball team` (0.91), `skateboarder` (0.89), and `golfer` (0.86) (results on Flickr30k, SD v2.0.), a phenomenon that has been previously

observed in VQA datasets [149]. Another observation is that *feminine* prompts also have a high association with food, such as `salad` (0.22), `meal` (0.25), and `cotton candy` (0.31) (results on Flickr30k, SD v2.0.). Additionally, results reveal that `businessman` (COCO, TextCaps, and Flickr30k on SD v1.4, COCO on SD v2.0, COCO and Flickr30k on SD v2.1) tends to be skewed towards *masculine* whereas `kitchenware` (GCC, SD v2.1) tends to be associated with *feminine* prompts.

3.6 Gender in Prompt-Image Dependencies

RQ3 *Does the gender in the input prompt influence the prompt-image dependencies in Stable Diffusion, and if so, which prompt-image dependencies are more predisposed to be affected?*

To answer this question, we need to know not only which objects are generated for each gender, but also how each object is generated in the diffusion process. To do so, we propose to classify objects into prompt-image dependency groups according to their relationship with the input prompt and the generated image. First, we conduct an *extended object extraction* by detecting not only the objects in the generated image, as in Section 3.5, but other objects also involved in the generative process. Then, we classify each object according to five *prompt-image dependency groups*, which allows us to study how gender influences objects according to their generative process.

3.6.1 Extended Object Extraction

To detect extended objects involved in the generative process, we conduct three extraction processes (see the example in "Prompt-image dependency" part of Figure 3.2).

(1) Nouns in prompt.

Prompts, designed by users, are a direct cue of what they wish to see in the generated image. The generated image, on the other hand, is required to be faithful to the prompt. The

first extraction process targets nouns within the prompt, recognizing their importance in directly shaping the occurrence of objects in the generated image. For each prompt, we obtain a noun set including all lemmatized nouns n in the prompt by using NLTK [150].

(2) Word attention.

Verifying whether objects in the noun set are faithfully generated in the image is demanding, as it requires locating the region that the noun guides. Fortunately, cross-attention has proven to be effective in exploring the word guidance during the generation process [105, 110]. Our second extraction process is the word attention masks generated by the cross-attention module via DAAM [110]. In detail, let $\mathbf{P}$ be a matrix whose column n is the word embedding corresponding to the word n in p, and $H(\mathbf{z}_t)$ be a feature map of a certain block of Stable Diffusion's UNet for latent embedding $\mathbf{z}_t$ in the t-th denoising step. Cross-attention between $\mathbf{P}$ and $H(\mathbf{z}_t)$ is given by

$$\mathbf{A}_t = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right), \tag{3.6}$$

where $\mathbf{Q}$ and $\mathbf{K}$ are the query and key matrices given using linear layers $\mathbf{W}_\text{Q}$ and $\mathbf{W}_\text{K}$ as $\mathbf{Q} = \mathbf{W}_\text{Q}H(\mathbf{z}_t)$ and $\mathbf{K} = \mathbf{W}_\text{K}\mathbf{P}$, whose output dimensionality is d (the index t for denoising step is omitted for simplicity). The heart of DAAM is $\mathbf{A}_t$, of which column n is the attention map from word n to each spatial position of feature map $H(\mathbf{z}_t)$. We aggregate the attention maps over UNet blocks, multiple attention heads, and denoising steps. Let α_n denote the attention map, reshaped and resized to the same size as the corresponding generated image x, normalized to $[0, 1]$. For each word, we first compute the normalized attention map, where a higher value indicates that the pixel is more associated with the word. Then, we binarize the attention map with a threshold θ to obtain a set of masks a_n, responding to the region of an object specified by the word n. In each prompt, we obtain a mask set containing the mask a_n for each word n. We set threshold θ as 0.35.

(3) Visual grounding.

Nouns and the corresponding object regions cover only a small subset of objects in the generated image; there should be many other objects that are not explicitly described in the

prompt, but are still included in the image to complete the scene. We aim to enumerate as many objects as possible for comprehensive object-level analysis. To spot regions of arbitrary objects, the last extraction process is the same visual grounding process as in Section 3.5.

3.6.2 Prompt-Image Dependency Groups

Next, we classify each detected object according to its generative process. On the one hand, the generated image should align with its prompt, which can be verified using the noun set and the mask set. On the other hand, the image may have other visual elements beyond the prompt, listed in the object set and the object region set. To define prompt-image dependency groups, we consider the dependency among objects, the noun set, and the mask set based on its membership.

Definition 3.6.1 (Explicitly)**.** If the object o is in the noun set, it is *explicitly* described in the prompt.

Definition 3.6.2 (Guided)**.** If object region m_o *sufficiently* overlaps with at least one mask in the mask set, the object o is *guided* by cross-attention between the prompt and the image. Sufficiency is determined by the coverage of object region m_o by the mask a,

$$\text{coverage}(m_o, a) = \frac{|m_o \cap a|}{|m_o|}, \tag{3.7}$$

where $|\cdot|$ is the number of pixels. Thus, if coverage(m_o, a) is larger than a certain threshold σ, the object region m_o sufficiently overlaps with the mask a. We set different values for threshold σ for words referring to humans (people, person, woman, women, man, men) versus objects. The reason is that for human words, the generated people in the images can still be considered as *guided* by those words, even when the coverage (Equation (3.7) of word attention and visual grounding is relatively low. For example, word attention on human words may focus only on the face, while the visual grounding covers the whole body. Since these partial overlap cases are common for human words, we set a lower threshold of $\sigma = 0.25$ when both the detected object and word refer to humans. For all other cases, a higher threshold of $\sigma = 0.7$ is used.

With these definitions, we cluster objects in the object set into five groups, as illustrated

in Figure 3.4 with the example prompt `young women having a picnic at the park during daytime`.

Explicitly guided. The object is *explicitly* mentioned in the prompt, and *guided* by cross-attention. Faithful image generation may require each noun to be associated with the corresponding object.

Implicitly guided. The object is *not explicitly* mentioned in the prompt, but *guided* by cross-attention. The object may be strongly associated with or pertain to a certain noun in the noun set, e.g., the object `basket` for the noun `picnic`.

Explicitly independent. The object is *explicitly* mentioned in the prompt, but *not guided* by cross-attention. e.g., `park`.

Implicitly independent. The object is *not explicitly* mentioned in the prompt, and *not guided* by cross-attention. The object is generated solely based on contextual cues, e.g., `grass`.

Hidden. The noun has no association with objects in the object set, i.e., the noun is *not included* in the images, e.g., `daytime`.

Figure 3.4 illustrates the object extraction processes and the resulting dependency groups. Dependency groups are important as they depict if an object tends to appear, for example, in relation to the prompt (*explicitly guided*) or just for filling the scene (*implicitly independent*). Together with the gender-specific sets of prompts, they vividly provide essential insights into how an image generation model behaves for different genders.

3.6.3 Result Analysis

We denote co-occurrence $C_g(o, \mathcal{P})$ as the number of occurrences of object o in each dependency group g. To clarify, given that there are nouns included in the hidden group, the computation of occurrence should be adjusted from $C_g(o, \mathcal{P})$ to $C_g(n, \mathcal{P})$ for n in the hidden group.

Objects in Dependency Groups

To answer RQ3, we first investigate objects in the prompt-image dependency groups, aiming to identify which types of objects are generated under the influence of the prompt, the cross-attention, or the context of the generated image. Figure 3.5 shows the prevalent objects within each dependency group across all datasets on SD v2.0 (to focus on the differences between generated objects, we remove individuals (person, people, women, woman, men, man, female, male, girl, boy)). Although the specific generated objects align with the prompt's domain, and their frequencies may vary across datasets, we observe consistent trends.

Objects in the *explicitly guided* group include animals and tangible items commonly encountered in daily life, such as `umbrella` and `table`. The *implicitly guided* group contains objects surrounding human beings, such as clothing and personal belongings like `shirt` and `goggles`. The *explicitly independent* group comprises words related to the surrounding environment, such as `kitchen` or `restaurant`. Objects in the *implicitly independent* group are typically part of the background that can be detected, like `tree` and `road`, along with attire accompanying individuals. Lastly, the *hidden* group comprises words challenging to detect in images, such as `game` and `air`.

Gender and Dependency Groups

Next, we investigate the relationship between gender and the objects in each prompt-image dependency group. To discern whether object differences are statistically significant, we conduct chi-square tests on the object co-occurrence for each dependency group. While we find significant differences (p-value < 0.05) across all datasets in the *implicitly guided* and *implicitly independent* groups, we do not find significant differences in most datasets in the *explicitly guided*, *explicitly independent*, and *hidden* groups. This suggests that while Stable Diffusion may consistently generate the nouns explicitly mentioned in the prompt, it may rely on gender cues for generating elements that are not specified in the prompt, such as the background and surroundings of the individuals.

To further explore the text-image dependencies and their correlation with gender, we calculate the bias score based on object co-occurrence in *implicitly guided* and the *implicitly independent* groups, both of which exhibit statistically significant differences. Figure 3.6 shows the top-10 objects skewed toward *masculine* and *feminine* in *implicitly guided* on all datasets and models. We analyze results on SD v1.4 as examples. For the *implicitly guided* group, we observe high bias scores for clothing items, such as `cocktail dress` (GCC, COCO, and Flickr30k(0.95)), `suit` (COCO(0.98), TextCaps(0.85)), and `bow tie` (GCC(0.98), COCO(0.95), TextCaps(0.91), Flickr30k and Profession(1)) for *masculine*, and `bikini` (GCC, COCO, and Profession(0), Flickr30k(0.04)), `dress` (COCO(0.12)) and `boot` (TextCaps(0.14)) for *feminine*, aligning with observations in a previous work [116]. Another prominent observation, consistent with the findings in RQ2, is the strong association of `child` (0.27) with *feminine*, and *masculine* with sports-related terms such as `player` (0.8) and `football player` (0.72) (results on TextCaps, SD v2.0.). Similar gendered associations are observed across different datasets and models.

Figure 3.7 shows the bias scores in *implicitly independent* across all datasets and models. While places and surroundings are the majority (as discussed in section 3.6.3), clothing associated with individuals in the *implicitly independent* group may exhibit higher or lower bias scores. Given the similar trend in clothing between *implicitly guided* and *implicitly independent*, we focus on surroundings and other items in the latter group. As the specific environments generated are influenced by the semantics of the text, we conduct analysis based on datasets. In COCO, results show that `basement` and `cabinet` are more prone to appear in *masculine*, while `dinner party`, and `passenger train` are inclined to be generated in *feminine*. In TextCaps, `grass`, `building`, and `field` are skewed toward *masculine*, while `park`, `carpet`, and `store` are skewed toward *feminine*. Taking GCC on SD v2.0 as an example, sports-related items such as `bodybuilder` (1) and `football team` (1) are again skewed toward *masculine*, while `instrument` (0.17) and `apron` (0.33) are more aligned with *feminine*. Additionally, there are also disparities related to backgrounds, such as `backdrop` (0.15) and `dirt field` (0.17) for *feminine*, and `stone building` (1) and

`tennis court` (0.63) for *masculine*. Furthermore, certain words consistently align with feminine across datasets and models, such as `smile` (TextCaps and Flickr30k on SD v1.4) and `flower` (COCO on SD v1.4).

3.7 Additional Experiments

To further evaluate our protocol, we conduct intra-prompt evaluation and human evaluation.

3.7.1 Intra-Prompt Evaluation

To eliminate the influence of randomness, we investigate the research questions using images generated from the same triplet prompts. We generate a total of 3000 images on 1000 seeds with SD v2.0, from triplet prompts derived from a caption in GCC: "*person looks at the falling balloons at the conclusion*". We use the same settings as conducted in the experiments above.

For RQ1, the representational disparities in Table 3.5 show that *neutral* is consistently closer to *masculine* in each space. For RQ2, the chi-square tests on the object occurrences among the triplets and every pair within the triplets, p-value is consistently less than 10^{-5}, indicating statistically significant differences. For RQ3, the chi-square tests also reveal significant differences in the groups *implicitly guided* and *implicitly independent* ($p < 10^{-5}$). However, we do not apply chi-square tests to *explicitly guided*, *explicitly independent*, and *hidden*, as the numbers of objects in these groups are less than 5. The co-occurrence similarity $s_{\mathrm{O}}(\mathcal{P}_{\mathrm{n}}, \mathcal{P}_{\mathrm{f}})$ between the neutral and feminine is 0.733, while the similarity $s_{\mathrm{O}}(\mathcal{P}_{\mathrm{n}}, \mathcal{P}_{\mathrm{m}})$ between the neutral and masculine is **0.773**. This indicates that the object co-occurrences in images generated from *neutral* prompts are closer to those from *masculine* prompts than *feminine* prompts. These findings correspond to the above results.

3.7.2 Dependency Groups Analysis

Taking Stable Diffusion v2.0 as an example, we scrutinize the dependency groups deeper to discover the underlying connections between groups and objects.

Dependency Group Presence in Images

To assess the presence of dependency groups in images, we compute the percentage of the images containing dependency groups over the total number of images in each dataset. The results are presented in Table 3.6. For example, in the GCC dataset, 64.48% images contain at least one object in the *explicitly guided* group. Similarly, in other datasets, over 60% of images have objects in the *explicitly guided*, except for the Profession set, where the proportion is only 15%. This disparity may be due to the specialized terminology in the Profession set, potentially reducing the chance of being detected by the visual grounding model. Conversely, only around 10% or fewer images include objects in the *explicitly independent* group. Given that most objects in this group represent the surrounding environment, objects in *explicitly independent* may occur when the prompt contains words indicating the surrounding environment (e.g., `park`, `kitchen`).

Moreover, a similar trend is observed across all datasets, where most images contain objects from the *implicitly guided*, *implicitly independent*, and *hidden* group. This indicates that text-to-image models generate auxiliary objects to fill in both the areas guided by the prompt and those independent from it. We posit that the high proportion of *hidden* group may be due to the abstract words that are challenging to detect and to the mismatch in synonyms. For instance, the visual grounding model may struggle to identify people as professions in the Profession set.

Amount of Objects

Next, we investigate the amount of individual objects in each dependency group and nouns in prompts. The results for each dataset are shown in Table 3.7. Supporting the findings in Table 3.6, objects in the *explicitly guided* and *explicitly independent* constitute only a small

portion of the nouns in the prompts. Additionally, despite not being mentioned in the prompt, *implicitly guided* and *implicitly independent* groups contain more objects than *explicitly* groups present in the image. This suggests that these two *implicitly* groups are worth further exploration for a comprehensive understanding of the image generation process.

3.7.3 Human Evaluation

To evaluate the reliability of the visual grounding model, we randomly select 100 generated images from SD v2.0 along with the nouns from the corresponding prompts and conduct a human evaluation to determine whether the nouns are present in the images. The 100 prompts contain 346 nouns, from which 227 (65.61%) are correctly identified both by humans and the automated vision grounding. Out of the remaining 119 nouns, only 8 nouns are detected by the model but not observed by humans. These nouns are `frisbee` (2), `women` (1), `people` (1), `kite` (1), `scooters` (1), `tennis` (1), and `speaker` (1). For the nouns not detected by the model but identified by humans, the most frequent ones are `woman` (10), `street` (7), `people` (6), and `snowy` (4). The absence of the noun `street` in the model's detection might be attributed to the strict alignment between nouns and objects. Even if the model successfully identifies `street scene`, the specific noun `street` might be placed in one of the *implicitly guided*, *implicitly independent*, or *hidden* groups. These results indicate that the visual grounding model has reasonable accuracy in detecting nouns appearing in the generated images, though there is still room for improvement on abstract nouns and scene-level nouns.

3.8 Recommendations

Our methodology revealed significant disparities in the objects generated by three Stable Diffusion models according to the gender in the input prompt. While these discrepancies may seem harmless, they can potentially reinforce gender stereotypes. With this in mind, we propose a series of suggested practices aimed at mitigating these concerns, both for model developers and for users:

3.8.1 Model Developers

Debias Text Embeddings

We have identified that gender bias originates in the text embedding, with *neutral* prompts consistently being more similar to *masculine* prompts than to *feminine* prompts, which propagates through the entire generation process. Given the documented presence of gender bias in CLIP [11, 151–153], it comes as no surprise that text-to-image generation models relying on CLIP also exhibit such biases. The first mitigation technique should focus on debiasing the text embedding space, aiming for more equitable representations.

Identify Problematic Representations

While some associations of certain objects with specific genders may not immediately raise concerns, others could potentially do so. Therefore, researchers must meticulously assess these associations, taking into account the cultural context in each instance. It is crucial to examine the co-occurrence of objects across genders and check whether neutral prompts tend to exhibit a preference toward a particular gender.

Investigate Modules That Complete the Scene

Significant differences were observed in the *implicitly* generated objects, underscoring the need to investigate how the model completes the scene. Future research could explore other modules, probing fine-grained control over the regions not guided by the input.

3.8.2 Users

Explicitly Specify Objects

Our results showed that there are no significant differences in the objects explicitly mentioned in the input prompts concerning gender. This suggests that Stable Diffusion models can adhere to the simple instructions in the prompt regardless of gender. Therefore, expanding the number

of objects in the input could offer greater control over broader guided regions and potentially lead to the generation of images with less gender disparity.

Explicitly Specify Gender

Considering that *neutral* prompts consistently produced images more similar to those from *masculine* prompts, we advise refraining from using neutral prompts if targeting a balanced distribution across genders. Instead, using prompts with specified gender indicators may be more reliable.

3.9 Limitations

We acknowledge that our proposed evaluation protocol has limitations, and we emphasize them here for transparency and to inspire the community to propose enhancements in future studies. Firstly, our evaluation protocol focuses on binary genders, neglecting to evaluate gender from a broader spectrum perspective. To enhance inclusivity, future research could extend the analysis to encompass a more diverse range of genders. Secondly, our protocol relies on a stringent alignment between nouns and objects, assuming their identity after lemmatization, which may overlook variations and synonyms. Thirdly, the objects segmented in visual grounding may encounter errors, possibly perpetuating issues in the classified groups. Additionally, if gender bias exists in the visual grounding model, where certain objects may be more challenging to detect in specific genders, this bias could transfer to the final results. Additionally, when the object comprises more than one word (e.g., “picnic basket”), each noun in the phrase has its own word attention rather than being considered as a single entity. Last but not least, our study only examines the presence of objects not differentiating with distinct attributes, such as color or shape.

Table 3.1: Gender bias evaluation methods in text-to-image generation. We compare with previous methods on input (prompt type, prompt variation), evaluation space (prompt, denoising, image), and bias (subject of bias). "Prompt variation" refers to how prompts vary in attributes (e.g., profession) while keeping other words unchanged when the prompts are template-based. If prompts are from caption datasets, the specific dataset names are presented. In terms of the "subject of bias", *gender* means the gender of generated faces, while *performance* contains generation performance metrics such as text-to-image alignment and image quality.

Method	Input		Evaluation Space			Bias
	Prompt Type	Prompt Variation	Prompt	Denoising	Image	Subject of Bias
[89]	Template	Identity, Profession	-	-	✓	Gender
[9]	Free-form	Objects	-	-	✓	Performance
[122]	Template	-	-	-	✓	Gender
[8]—Fairness	Free-form	COCO [137]	-	-	✓	Performance
[8]—Bias	Template	Adjective, Profession	-	-	✓	Gender
[92]	Template	Profession	-	-	✓	Gender, Attire
[10]	Template	Profession	-	-	✓	Gender
[93]—Profession	Template	Profession	✓	-	✓	Gender
[93]—Science/Career	Template	Science, Career	-	-	✓	Gender
[123]	Free-form	Creative prompts, Diffusion DB [138]	-	-	✓	Concept
[116]	Template	Attire, Activity	-	-	✓	Attire
[95]	Template	Adjective, Profession	-	-	✓	Gender
[95]—Expanded	Template	Profession	-	-	✓	Gender, Performance
[124]	Template	Adjective, Profession, Multilingualism	-	-	✓	Gender
[125]	Free-form	Profession	-	-	✓	Gender
[126]	Template	Profession, Social relation, Adjective	-	-	✓	Gender
[127]	Template	Action, Appearances	-	-	✓	Gender
[128]	Template	Two professions	-	-	✓	Gender
[129]	Template	Activity, Object, Adjective, Profession	-	-	✓	Gender
[130]	Free-form	Flickr30k [139], COCO [137]	-	-	✓	Gender
[115]	Template	-	-	-	✓	Object
[87]	Free-form	PHASE [87]	-	-	✓	Safety
[131]	Template	Profession, Sports, Objects, Scene	✓	-	✓	Gender
Ours	Free-form	GCC [140], COCO [137], TextCaps [141], Flickr30k [139], Profession	✓	✓	✓	Layout, Objects

Type	Word
Gender	woman, female, lady, mother, girl, aunt, wife, actress, princess, waitress, sister, queen, pregnant, daughter, she, her, hers, herself, bride, mom, queen, man, male, father, gentleman, boy, uncle, husband, actor, prince, waiter, son, brother, guy, emperor, dude, cowboy, he, his, him, himself, groom, dad, king
Geography	American, Asian, African, Indian, Latino
Others	commander, officer, cheerleader, couple, player, magician, model, entertainer, astronaut, artist, student, politician, family, guest, driver, friend, journalist, relative, hunter, tourist, chief, staff, soldier, civilian, author, prayer, pitcher, singer, kid, groomsman, bridemaid, ceo, customer, dancer, photographer, teenage, child, u, me, I, leader, crew, athlete, celebrity, priest, designer, hiker, footballer, hero, victim, manager, Mr, member, partner, myself, writer

Table 3.2: Words that indicate humans.

Table 3.3: Number of generated triplets, prompts, and images for each dataset.

Data	Triplets	Prompts	Seeds	Images
GCC (val)	418	1254	5	6270
COCO	51,219	153,657	1	153,657
TextCaps	4041	12,123	1	12,123
Flickr30k	16,507	49,521	1	49,521
Profession	811	2433	5	12,165

Table 3.4: Co-occurrence similarity on Stable Diffusion models.

Pairs	GCC	COCO	TextCaps	Flickr30k	Profession
SD v1.4					
$s_{\mathrm{O}}(\mathcal{P}_{\mathrm{n}}, \mathcal{P}_{\mathrm{f}})$	0.379	0.486	0.413	0.424	0.350
$s_{\mathrm{O}}(\mathcal{P}_{\mathrm{n}}, \mathcal{P}_{\mathrm{m}})$	**0.414**	**0.516**	**0.444**	**0.457**	**0.374**
SD v2.0					
$s_{\mathrm{O}}(\mathcal{P}_{\mathrm{n}}, \mathcal{P}_{\mathrm{f}})$	0.382	0.512	0.420	0.445	0.362
$s_{\mathrm{O}}(\mathcal{P}_{\mathrm{n}}, \mathcal{P}_{\mathrm{m}})$	**0.425**	**0.531**	**0.448**	**0.476**	**0.376**
SD v2.1					
$s_{\mathrm{O}}(\mathcal{P}_{\mathrm{n}}, \mathcal{P}_{\mathrm{f}})$	0.380	0.499	0.388	0.426	0.349
$s_{\mathrm{O}}(\mathcal{P}_{\mathrm{n}}, \mathcal{P}_{\mathrm{m}})$	**0.419**	**0.522**	**0.419**	**0.451**	**0.382**

Table 3.5: Representational disparities between the neutral, feminine, and masculine in the three spaces from intra-prompts (SD v2.0).

Pairs	Prompt	Denoising	Image					
	t	z_0	*SSIM* ↑	*Diff. Pix.* ↓	*ResNet* ↑	*CLIP* ↑	*DINO* ↑	*Split-Product* ↑
(neu, fem)	0.981	0.789	0.547	37.54	0.867	0.844	0.557	0.947
(neu, mas)	**0.982**	**0.829**	**0.587**	**33.81**	**0.892**	**0.864**	**0.625**	**0.959**

Table 3.6: The proportion of images containing the dependency groups to all the images for each dataset on SD v2.0.

Dataset	Explicitly Guided	Implicitly Guided	Explicitly Independent	Implicitly Independent	Hidden
GCC	64.48	90.70	7.81	59.11	96.14
COCO	83.67	93.54	10.47	57.53	92.61
TextCaps	61.97	86.60	8.78	61.90	99.10
Flickr30k	83.07	94.91	9.56	58.89	92.48
Profession	15.03	98.07	3.48	63.22	100.00

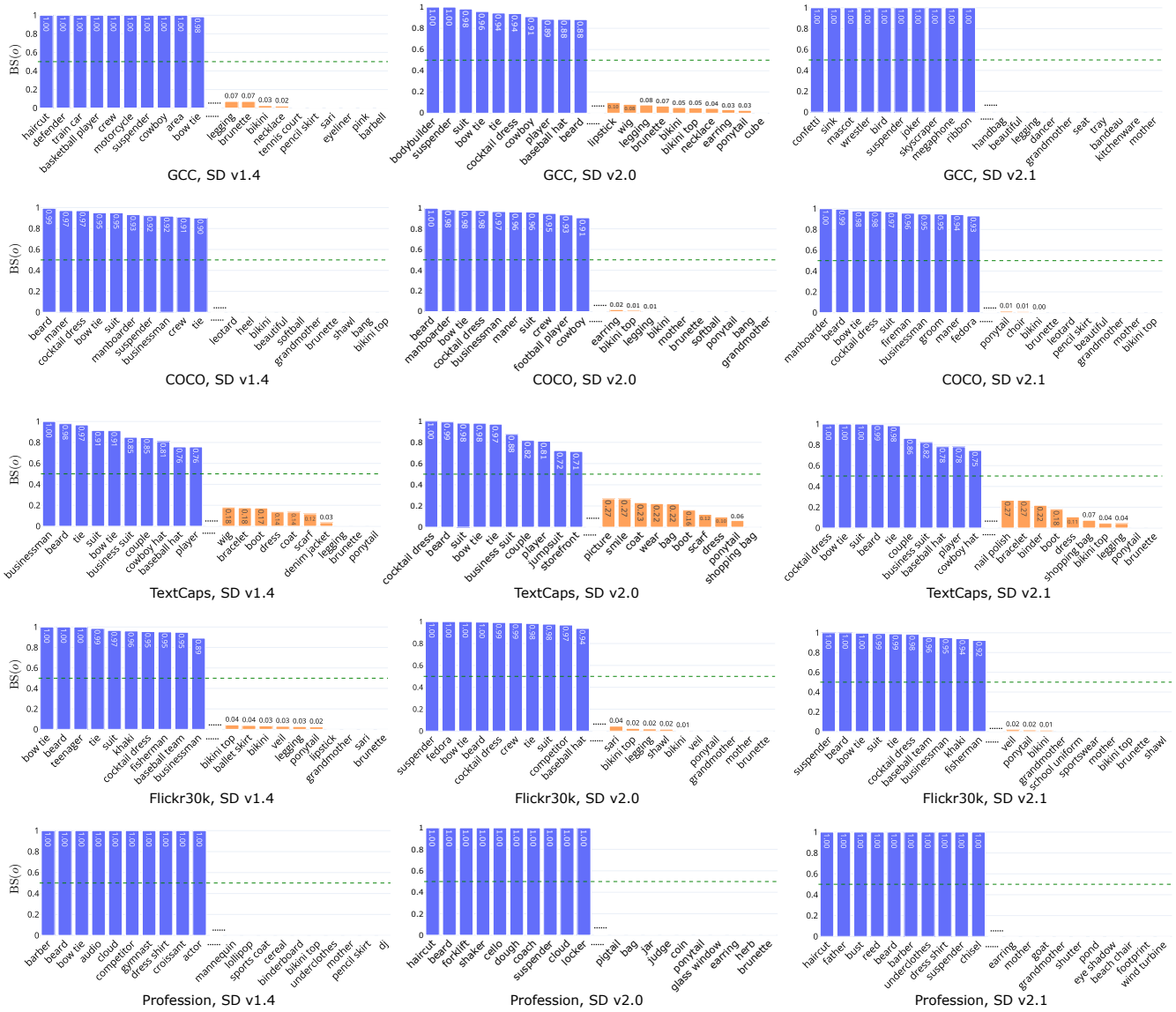

Figure 3.3: Bias score on all datasets (rows) and models (columns). A high score (blue) indicates the object appears more frequently in masculine, while a low score (orange) suggests the object is more commonly shown in feminine.

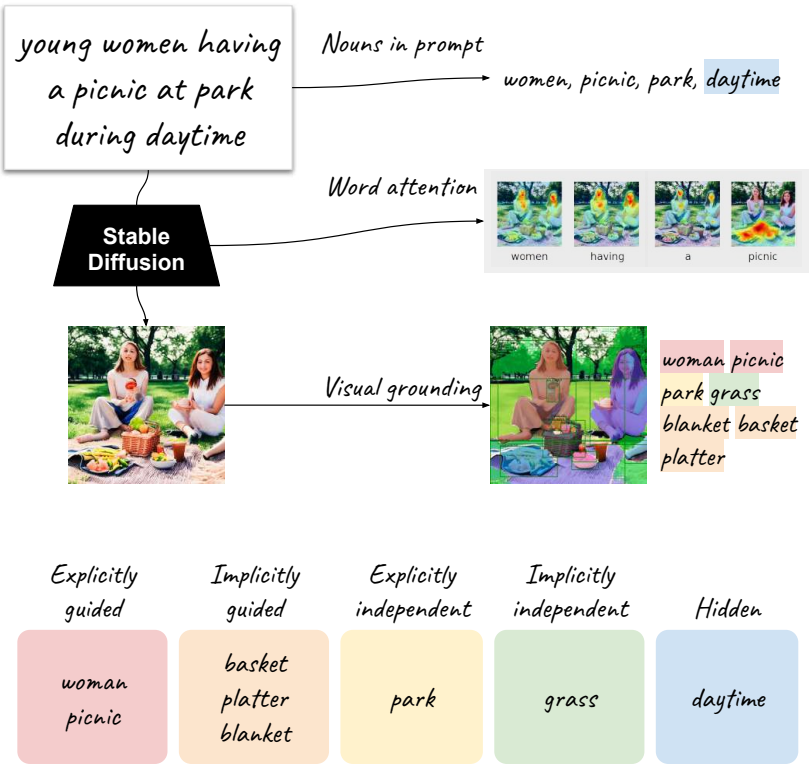

Figure 3.4: Prompt-image dependency groups.

Table 3.7: Amount of individual objects in each dependency group and nouns in prompts on SD v2.0 for each dataset.

Dataset	Explicitly Guided	Implicitly Guided	Explicitly Independent	Implicitly Independent	Hidden	Nouns
GCC	155	1059	85	625	536	544
COCO	827	2418	391	1529	3274	3305
TextCaps	371	1347	147	741	3608	3638
Flickr30k	659	2017	330	1255	2718	2741
Profession	162	1331	76	650	1041	1043

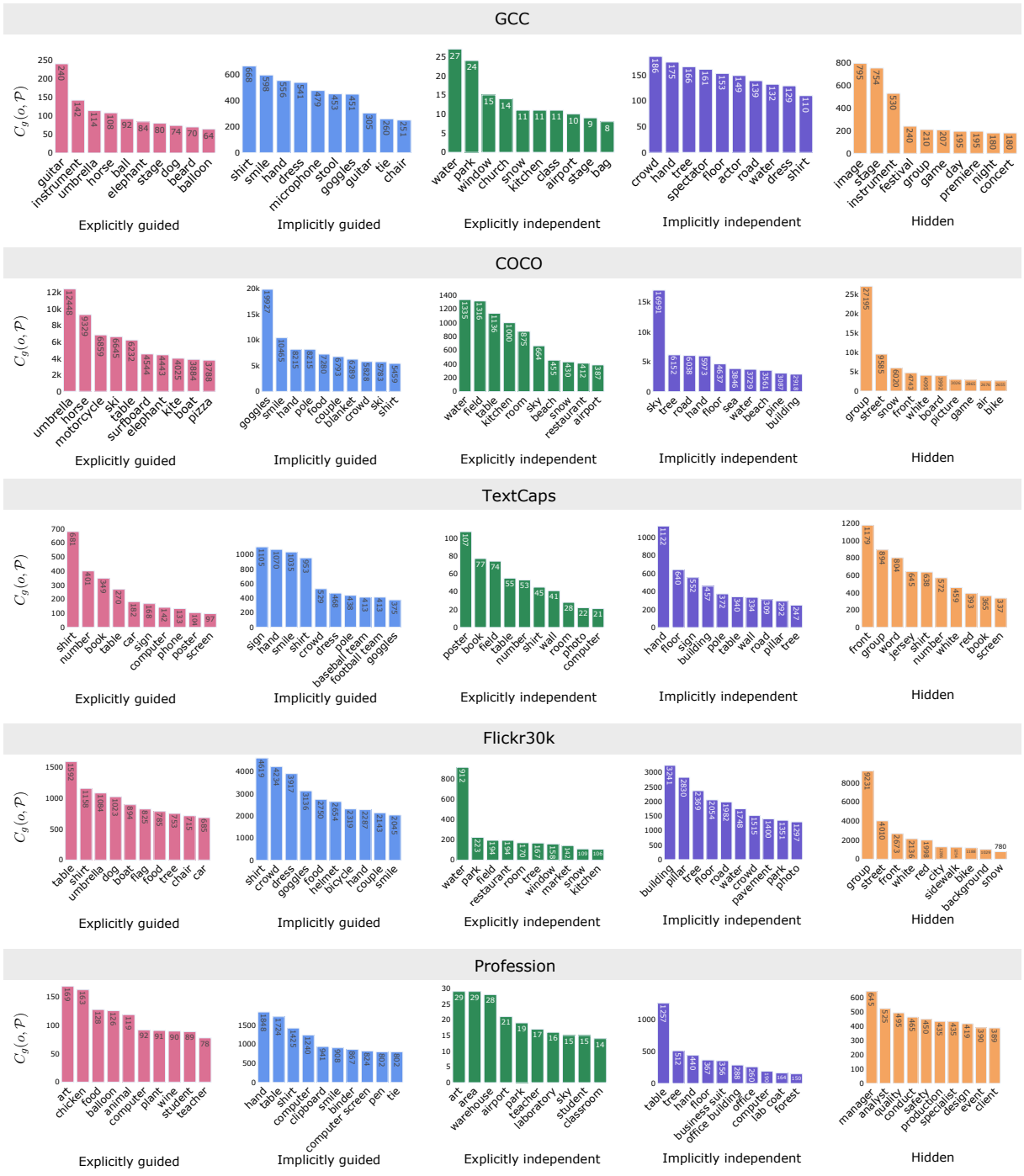

Figure 3.5: The occurrence $C_g(o, \mathcal{P})$ of object o in images generated from $\mathcal{P}$ on each dependency group for each dataset (SD v2.0).

Figure 3.6: Bias score on *implicitly guided* on the datasets (rows) and models (columns).

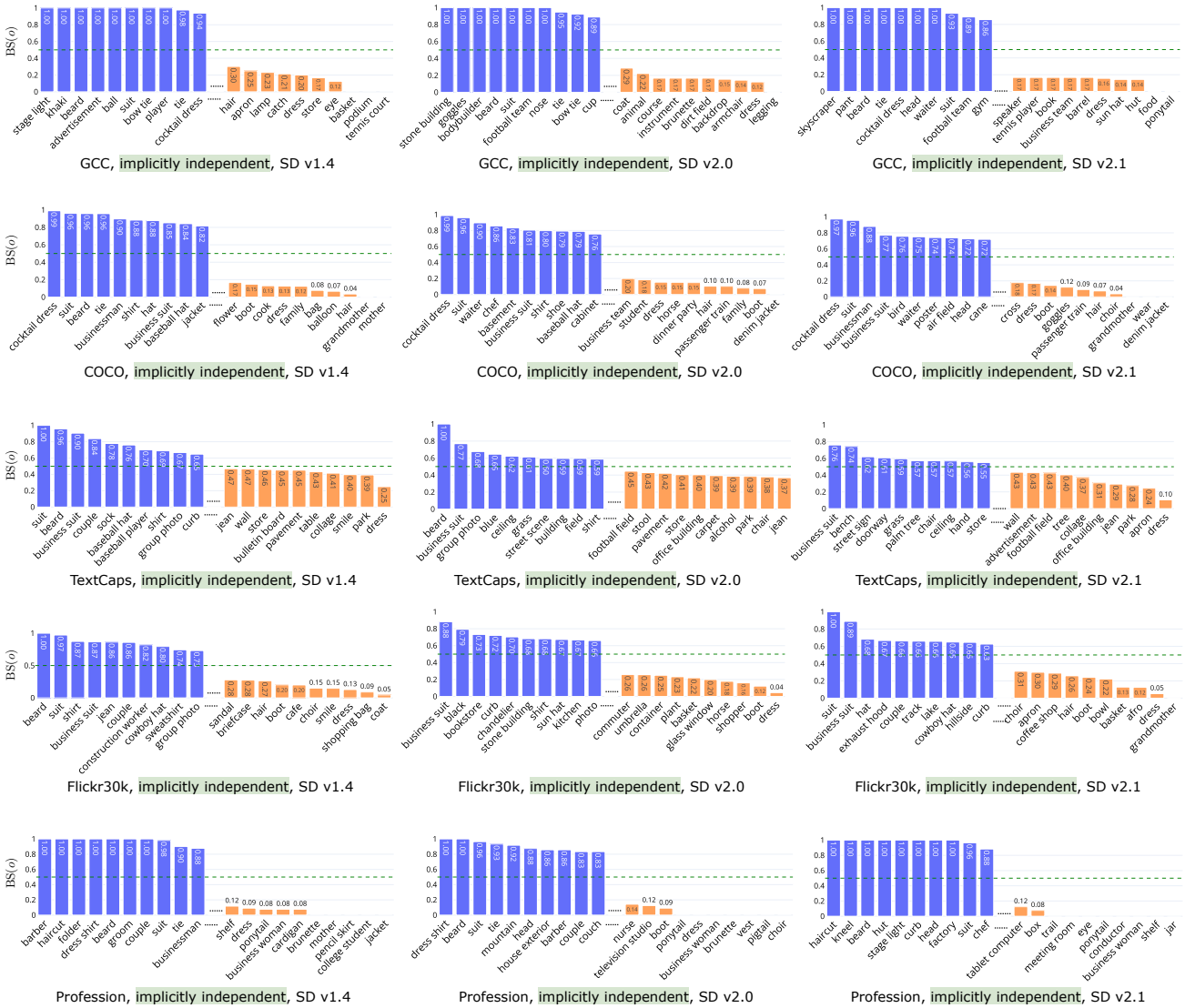

Figure 3.7: Bias score on *implicitly independent* on the datasets (rows) and models (columns).

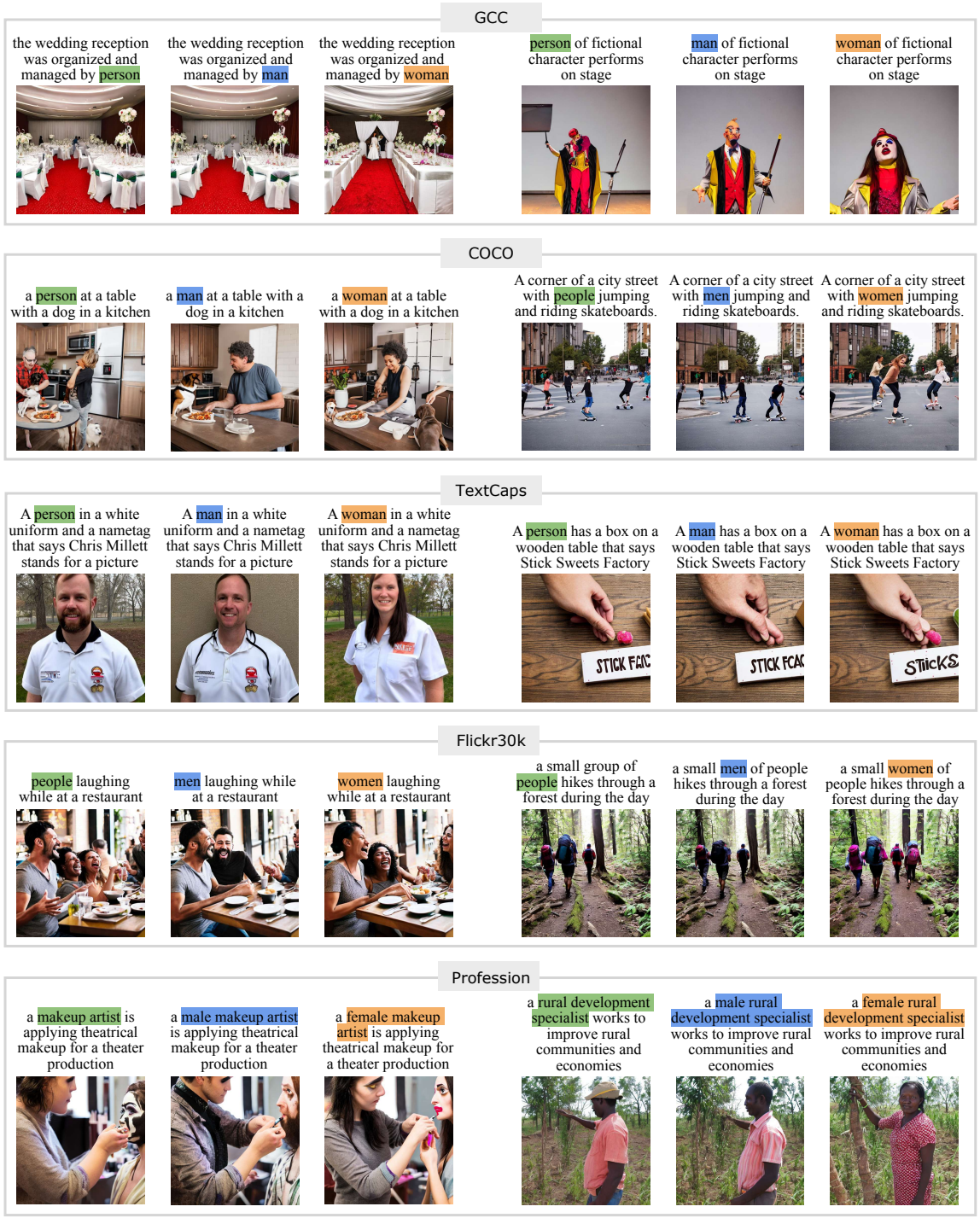

Figure 3.8: Examples of triplet prompts and the corresponding generated images for each dataset on SD v2.0.

Table 3.8: Representational disparities between neutral, feminine, and masculine prompts in the three spaces on Stable Diffusion models.

Pairs	Prompt	Denoising	Image					
	t	z_0	*SSIM* ↑	*Diff. Pix.* ↓	*ResNet* ↑	*CLIP* ↑	*DINO* ↑	*Split-Product* ↑
SD v1.4								
GCC								
(neu, fem)	0.909	0.770	0.516	42.61	0.848	0.794	0.543	0.956
(neu, mas)	**0.931**	**0.798**	**0.543**	**39.34**	**0.859**	**0.808**	**0.576**	**0.961**
COCO								
(neu, fem)	0.920	0.778	0.568	38.558	0.866	**0.8584**	0.564	0.957
(neu, mas)	**0.942**	**0.796**	**0.592**	**35.671**	**0.873**	0.8580	**0.591**	**0.959**
TextCaps								
(neu, fem)	0.931	0.747	0.461	46.873	0.853	0.773	0.530	0.952
(neu, mas)	**0.948**	**0.768**	**0.487**	**43.599**	**0.862**	**0.786**	**0.555**	**0.954**
Flickr30k								
(neu, fem)	0.913	0.792	0.492	44.010	0.858	**0.830**	0.563	0.959
(neu, mas)	**0.931**	**0.804**	**0.518**	**41.105**	**0.865**	0.828	**0.587**	**0.960**
Profession								
(neu, fem)	0.854	0.765	0.487	45.006	0.831	0.830	0.528	0.948
(neu, mas)	**0.862**	**0.783**	**0.508**	**42.528**	**0.843**	**0.846**	**0.555**	**0.952**
SD v2.0								
GCC								
(neu, fem)	0.980	0.767	0.543	39.00	0.847	0.797	0.545	0.957
(neu, mas)	**0.982**	**0.790**	**0.571**	**35.82**	**0.864**	**0.817**	**0.581**	**0.963**
COCO								
(neu, fem)	0.984	0.793	0.603	34.10	0.881	**0.861**	0.595	0.9645
(neu, mas)	**0.985**	**0.805**	**0.616**	**32.50**	**0.887**	0.859	**0.609**	**0.9647**
TextCaps								
(neu, fem)	0.9846	0.745	0.502	41.41	0.861	0.771	0.536	0.958
(neu, mas)	**0.9854**	**0.767**	**0.530**	**37.41**	**0.874**	**0.791**	**0.570**	**0.962**
Flickr30k								
(neu, fem)	0.982	0.801	0.541	38.42	0.871	**0.833**	0.584	0.9685
(neu, mas)	**0.983**	**0.809**	**0.559**	**36.02**	**0.874**	0.826	**0.601**	**0.9686**
Profession								
(neu, fem)	**0.85784**	0.766	0.511	42.41	0.839	0.846	0.537	0.952
(neu, mas)	0.85783	**0.779**	**0.528**	**40.71**	**0.848**	**0.857**	**0.556**	**0.953**
SD v2.1								
GCC								
(neu, fem)	0.980	0.755	0.522	41.48	0.842	0.805	0.527	0.952
(neu, mas)	**0.982**	**0.782**	**0.552**	**37.96**	**0.856**	**0.820**	**0.566**	**0.959**
COCO								
(neu, fem)	0.984	0.763	0.569	37.796	0.8670	**0.858**	0.556	0.955
(neu, mas)	**0.985**	**0.780**	**0.586**	**35.632**	**0.8747**	0.853	**0.575**	**0.957**
TextCaps								
(neu, fem)	0.9846	0.713	0.456	46.600	0.838	0.752	0.492	0.948
(neu, mas)	**0.9854**	**0.747**	**0.483**	**43.362**	**0.851**	**0.773**	**0.524**	**0.953**
Flickr30k								
(neu, fem)	0.982	0.772	0.499	42.722	0.853	**0.823**	0.544	**0.9572**
(neu, mas)	**0.983**	**0.784**	**0.511**	**40.988**	**0.857**	0.813	**0.555**	0.9570
Profession								
(neu, fem)	**0.85784**	0.759	0.497	44.173	0.835	0.856	0.521	0.945
(neu, mas)	0.85783	**0.778**	**0.517**	**41.796**	**0.848**	**0.870**	**0.548**	**0.947**

Chapter 4

Mitigating Gender Bias on Stable Diffusion

4.1 Overview

Text-to-image generation has demonstrated remarkable capabilities in generating high-quality images conditioned on natural language prompts. However, ethical concerns such as fairness and bias have gained increasing attention [10, 89, 90]. As revealed in Chapter 3, the inherent gender bias cannot be neglected – the neutral prompts (*e.g.*, "a person is playing basketball") tend to produce images that are more visually and semantically aligned with those generated from masculine prompts (*e.g.*, "a man is playing basketball") than from the feminine ones (*e.g.*, "a woman is playing basketball"). This preference not only reflects underlying societal biases in the output images but also recovers bias embedded in the latent representations within these generative models.

To address bias in generative models, existing work has primarily focused on facial attributes (*e.g.*, change gender cues such as hairstyle, facial structure, makeup, or other attributes like glasses) [118, 154]. However, this narrow focus overlooks other important elements in the images that may also contain gender bias implicitly, such as outfits and background, as shown

in Chapter 3, and features like object color or contextual cues that are hard to detect by computer vision tools. Moreover, most of the existing approaches rely heavily on fine-tuning the foundation model, which is computationally expensive and potentially harmful to the model's performance on unrelated tasks or domains [155].

To address this issue, we introduce a novel, training-free method to mitigate gender bias in the entire generated images without the need for any fine-tuning process. Our approach operates within the existing latent spaces of the model, allowing for a lightweight and plug-and-play solution without altering the model weights while providing fairer and more diverse outputs. Specifically, we aim to generate images that are both neutral and diverse by leveraging a designed interpolation strategy. This strategy constructs new and fair representations by interpolating between the feminine and masculine counterparts, enabling the generation of outputs that better reflect a neutral concept. Our method is implemented using MM-DiT, the backbone of Stable Diffusion 3 [1], one of the cutting-edge models.

The core of our method lies in an interpolation framework that operates across both the text space and the attention mechanism. Given a neutral input prompt (*e.g.*, "a person is playing basketball"), we first construct the feminine and masculine prompts by replacing the neutral term *person* with the gendered indicator (*e.g.*, *woman* and *man*), respectively. These counterpart prompts are encoded by the text encoder in the model to generate the corresponding text embeddings. Since bias originates from text embedding, as indicated in Chapter 3, we first form a debiased text embedding by interpolating the text embeddings from feminine and masculine prompts, aiming to start with the intended neutrality of the input condition. Beyond text space, we extend our interpolation mechanism into the attention modules in the model. During the iterative denoising process, we apply interpolation on the attention outputs and effectively blend the feminine and masculine. The interpolation in both semantic and structural components enables the model to explore sampling from the latent space, guiding the generation toward genuinely neutral content.

To regulate the interpolation trajectory, we employ a Beta prior over the interpolation coefficient, which is updated dynamically throughout the generation process, following [156].

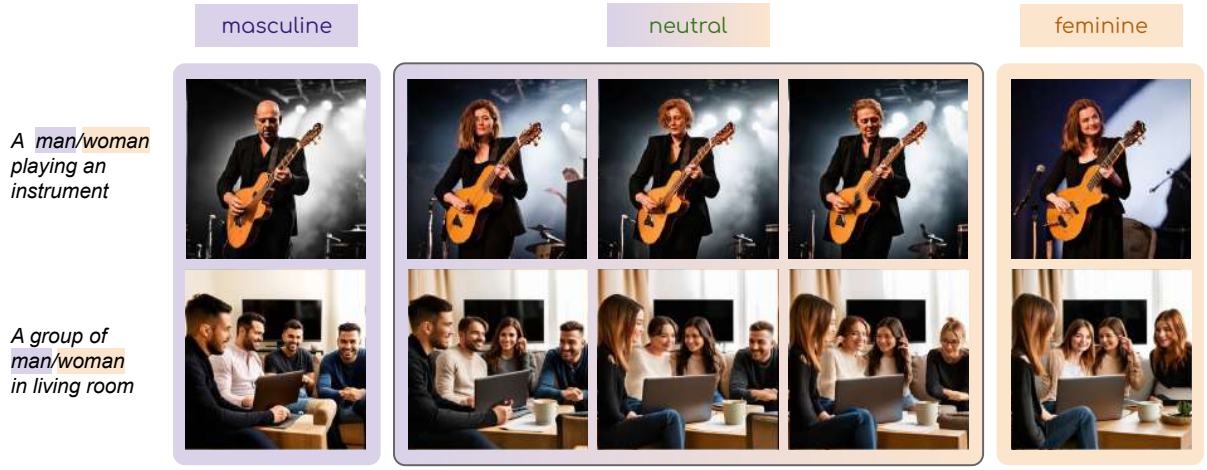

Figure 4.1: Examples of interpolated images between masculine and feminine outputs. The masculine and feminine images are generated by Stable Diffusion 3 [1] using the gendered prompt shown on the left. The three images in the middle represent interpolations. In the first row, the interpolated images exhibit diverse facial attributes. In the second row, the method maintains high visual quality even when generating images of multiple people, while also varying facial features effectively.

This distribution allows us to efficiently choose the parameters that concentrate in a small range under specific α and β values, encouraging valid interpolation effectively. After generating a sequence of candidate interpolation images, we compute perceptual distances among them to identify the smoothest and most coherent path, encouraging continuity in the interpolation process. From this path, one image is randomly selected to represent the final *neutral* output. The selected image is not a simple midpoint but rather a semantically fair outcome, preserving the original prompt semantically while mitigating gender bias.

Our experimental evaluation on bias demonstrates that the method effectively reduces the similarity disparities between the neutral and gendered images, providing more balanced outputs for neutral. Additionally, the evaluation on consistency, smoothness, and image fidelity also exhibits the strength of our method in effectively reducing gender bias while maintaining high image quality. Notably, our method performs well in mitigating bias in the entire image

across a variety of prompts from vision-language datasets. Furthermore, as our method does not rely on fine-tuning the model or additional data, it can be easily applied as a post-processing improvement to any pre-trained Stable Diffusion 3 model or similar architecture. Overall, this chapter introduces a practical and effective strategy for gender bias mitigation in text-to-image models that avoids the complexities of fine-tuning. By applying interpolations in both text embedding and attention modules, we propose a training-free method that facilitates diverse and fair image generation given neutral prompts, offering a lightweight and practical technique to mitigate gender bias in one of the most influential text-to-image generation models.

4.2 Related work

4.2.1 Image editing

Recent advances in text-to-image diffusion models have underscored the importance of controllability, ensuring that generated images faithfully reflect the semantics of complex prompts. Most of the work heavily relies on attention manipulation techniques as the cross-attention maps capture the alignment between textual tokens to visual regions [105, 157–162]. For example, Prompt-to-Prompt [105] enables localized editing through attention reweighting, while A-STAR [158] introduces attention segregation and retention losses to mitigate overlap and object missing. These methods are training-free and operate by injecting linguistic or structural priors into attention modules, improving attribute binding and compositional correctness without altering model weights. However, these methods are designed for diffusion models with a UNet [142] backbone, such as Stable Diffusion v1/v2/XL, whereas recent state-of-the-art diffusion models (*e.g.*, Stable Diffusion 3 [1]) use MM-DiT, which employ transformer-based fused attention modules, where the manipulation strategies remain underexplored.

Besides the cross attention module, manipulation in the latent or text embedding space provides a complementary way for image editing [154, 156, 163, 164]. NAO [163] uses norm-aware latent interpolation to improve the generation of rare concept images. Methods such

as AID [156] and recent work on singular value decomposition(SVD)-based embedding control [164] demonstrate that text embeddings encode semantically editable operations like object substitution and style transfer. These strategies on the embedding level are often paired with guidance from CLIP, reaching a balance between precision and flexibility [154, 156, 165]. Our proposed method builds upon these directions, offering a unified and training-free mechanism that combines attention and embedding interpolation to mitigate gender bias in image generation based on the latest MM-DiT structure.

4.2.2 Bias mitigation in text-to-image generation

Many prior studies have aimed to mitigate social biases in text-to-image generation, particularly those related to gender, race, and occupation [161, 166, 167]. A common strategy is to improve at the model level, including fine-tuning model on curated data [121, 166–168], editing specific model modules [109, 161, 162, 169] or introducing external constraints or conditions [118, 170–172] to control the generation toward more balanced outputs.

Training-free strategies applied on the inference stage have also gained attention due to their efficiency and ease of deployment. Some work mitigates bias at the level of the prompts or text embeddings [173, 174]. For example, Chuang *et al.* [173] mitigate bias by projecting out the biased directions in the text embeddings, while PreciseDebias [174] uses fine-tuned LLMs to rewrite generic prompts into demographically-informed ones aligned with specified distributions. Another widely explored focus involves modifying the guidance direction to steer the model toward a desired direction [154, 175–180]. Beyond prompt and guidance manipulation, the attention mechanism plays a critical role in mitigating bias during the denoising process [161, 169, 181]. For example, MIST [169] uses [EOS] embeddings to disentangle attributes and update cross-attention to address intersectional bias. Similarly, linguistically aligned attention guidance [161] adjusts attention weights based on syntactic relations to achieve fairer image-text alignment. These methods highlight the importance of the attention module in shaping fair generation outcomes.

Building on these insights, our work proposes a training-free debiasing strategy that manipulates both text embedding and attention interpolation within the MM-DiT architecture. Unlike prior research that focuses only on a single module or shifts the model toward a certain direction to one side (*e.g.*, either female or male), we produce genuinely *neutral* generations that are not biased toward any specific gender. By interpolating across semantic and attention spaces without conditioning on predefined attributes (*e.g.*, enforcing a 50 to 50 male-female split), our method enables the generation of diverse and inclusive outputs while maintaining semantic consistency and visual fidelity. Our method is plug-and-play, offering an efficient and practical solution for fairness in text-to-image generation.

4.3 Preliminary

4.3.1 Text-to-image generation

Diffusion-based models have become the foundation of high-quality text-to-image generation [1, 3, 4]. Stable Diffusion 3 (SD3) adopts a rectified flow-based approach that gradually transforms Gaussian noise into a high-resolution image guided by a text prompt. Unlike traditional diffusion models that iteratively denoise with stochastic noise schedules, SD3 formulates the generative process as a continuous flow in the latent space, mapping noise directly to data along a straight path. Specifically, the model operates on an image embedding x_0, and learns to transform a Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ to x_0 through a learned velocity field. The forward trajectory can be defined as a linear interpolation between data and noise:

$$\mathbf{z}_T = (1 - T)x_0 + T\epsilon, \quad T \in [0, 1]. \tag{4.1}$$

where $\mathbf{z}_T$ is the intermediate latent state at time T, and the model is trained to predict the velocity of this straight path:

$$v_{target} = \epsilon - x_0. \tag{4.2}$$

To approximate this velocity field, SD3 uses a multimodal transformer architecture (MM-DiT), which jointly encodes the noise latent $\mathbf{z}_T$ at time T, and a text embedding c derived from the concatenation of three frozen text encoders. The network outputs the predicted velocity vector $v_\theta(\mathbf{z}_T, T, c)$. The training phase minimizes the following conditional flow matching loss:

$$\mathcal{L} = \mathbb{E}_{x_0,\epsilon,T}\left[\left\|v_\theta\left(\mathbf{z}_T, T, c\right) - (\epsilon - x_0)\right\|^2\right], \tag{4.3}$$

which encourages the network to predict the correct directional flow in the latent space that aligns noise with the ground-truth image under the condition of the text prompt.

At inference time, the model starts from a Gaussian noise sample and integrates the learned velocity field backward from $T = 1$ to $T = 0$ to predict the denoised latent $\mathbf{z}_0$, which is subsequently decoded into the final image. This approach enables SD3 to generate semantically aligned and high-fidelity images with fewer sampling steps compared to traditional diffusion models.

4.3.2 Fused attention in Stable Diffusion 3

Stable Diffusion 3 proposes MM-DiT, a transformer-based architecture that integrates visual and textual inputs through a fused attention mechanism. Rather than using cross-attention to connect text and image features, SD3 treats image and text embeddings symmetrically within a unified attention block. This setup allows both visual and textual features to interact with each other, leading to improved faithful generation.

Let $\mathbf{z}_T \in \mathbb{R}^{n_i \times d}$ denote the image embedding at the timestep T, and $c \in \mathbb{R}^{n_t \times d}$ represent the text embeddings. These are independently projected to generate their respective queries, keys, and values:

$$Q_i = W_i^Q \mathbf{z}_T, \quad K_i = W_i^K \mathbf{z}_T, \quad V_i = W_i^V \mathbf{z}_T, \tag{4.4}$$

$$Q_t = W_t^Q c, \quad K_t = W_t^K c, \quad V_t = W_t^V c, \tag{4.5}$$

where $W_i^Q, W_i^K, W_i^V, W_t^Q, W_t^K, W_t^V \in \mathbb{R}^{d \times d_k}$ [1] are learnable projection matrices, and d_k is the head dimension. The fused attention computation is then applied to concatenated queries, keys, and values:

$$Q = \begin{bmatrix} Q_i \\ Q_t \end{bmatrix}, \quad K = \begin{bmatrix} K_i \\ K_t \end{bmatrix}, \quad V = \begin{bmatrix} V_i \\ V_t \end{bmatrix}. \tag{4.6}$$

The full attention output is computed using standard scaled dot-product attention:

$$\begin{aligned}
\text{Attn}(Q, K, V) &= \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \\
&= \text{softmax}\left(\frac{\begin{bmatrix} Q_i \\ Q_t \end{bmatrix} \begin{bmatrix} K_i^\top, K_t^\top \end{bmatrix}}{\sqrt{d_k}}\right) \begin{bmatrix} V_i \\ V_t \end{bmatrix} \\
&= \begin{bmatrix} \text{softmax}\left(\frac{Q_i[K_i^\top, K_t^\top]}{\sqrt{d_k}}\right) \\ \text{softmax}\left(\frac{Q_t[K_i^\top, K_t^\top]}{\sqrt{d_k}}\right) \end{bmatrix} \begin{bmatrix} V_i \\ V_t \end{bmatrix} \\
&= \begin{bmatrix} \text{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_k}}\right) & \text{softmax}\left(\frac{Q_i K_t^\top}{\sqrt{d_k}}\right) \\ \text{softmax}\left(\frac{Q_t K_i^\top}{\sqrt{d_k}}\right) & \text{softmax}\left(\frac{Q_t K_t^\top}{\sqrt{d_k}}\right) \end{bmatrix} \begin{bmatrix} V_i \\ V_t \end{bmatrix} \\
&= \begin{bmatrix} \text{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_k}}\right) V_i + \text{softmax}\left(\frac{Q_i K_t^\top}{\sqrt{d_k}}\right) V_t \\ \text{softmax}\left(\frac{Q_t K_i^\top}{\sqrt{d_k}}\right) V_i + \text{softmax}\left(\frac{Q_t K_t^\top}{\sqrt{d_k}}\right) V_t \end{bmatrix} \\
&= \begin{bmatrix} \text{Attn}(Q_i, K_i, V_i) + \text{Attn}(Q_i, K_t, V_t) \\ \text{Attn}(Q_t, K_i, V_i) + \text{Attn}(Q_t, K_t, V_t) \end{bmatrix}
\end{aligned} \tag{4.7}$$

Here, $\text{Attn}(Q_i, K_i, V_i)$ denotes the self attention from image to image, $\text{Attn}(Q_i, K_t, V_t)$ denotes cross attention from image to text, $\text{Attn}(Q_t, K_i, V_i)$ denotes reversed cross attention from text to image, and $\text{Attn}(Q_t, K_t, V_t)$ denotes self attention from text to text.

This decomposition illustrates that each query distributes attention over all key-value pairs, and the final output is formed by additive contributions from each modality. Not only do image

[1] The timestep T is omitted for simplicity.

queries interact with both image and text features, but text queries also attend to image content. This enables fine-grained semantic alignment and compositional generation. By unifying intra-modal and inter-modal interactions within a single attention layer, MM-DiT simplifies the architecture while enhancing its multimodal expressiveness.

4.3.3 Triplet generation

To evaluate gender bias and apply bias mitigation, we construct triplet prompts following the method described in Chapter 3. In detail, we use captions sourced from four vision-language datasets (GCC validation set [140], COCO [137], TextCaps [141], and Flickr30k [139]), as well as a sentence set Profession generated from ChatGPT 3.5 [12]. From these captions, we choose sentences that contain the word *person* or *people* as neutral prompts. We then create feminine and masculine counterpart prompts by replacing *person/people* with *woman/women* and *man/men*, respectively. For the profession set, we prepend *female/male* to the profession name to construct feminine and masculine prompts.

Let p_{n}, p_{f}, and p_{m} present the neutral, feminine, and masculine prompt, respectively. Their corresponding text embeddings are denoted as c_{n}, c_{f}, and c_{m}. Within the MM-DiT architecture of Stable Diffusion 3, the fused attention output for each prompt type can be presented using the following decompositions:

$$\mathrm{Attn}(Q^{\mathrm{n}}, K^{\mathrm{n}}, V^{\mathrm{n}}) = \begin{bmatrix} \mathrm{Attn}(Q_i^{\mathrm{n}}, K_i^{\mathrm{n}}, V_i^{\mathrm{n}}) + \mathrm{Attn}(Q_i^{\mathrm{n}}, K_t^{\mathrm{n}}, V_t^{\mathrm{n}}) \\ \mathrm{Attn}(Q_t^{\mathrm{n}}, K_i^{\mathrm{n}}, V_i^{\mathrm{n}}) + \mathrm{Attn}(Q_t^{\mathrm{n}}, K_t^{\mathrm{n}}, V_t^{\mathrm{n}}) \end{bmatrix} \tag{4.8}$$

$$\mathrm{Attn}(Q^{\mathrm{f}}, K^{\mathrm{f}}, V^{\mathrm{f}}) = \begin{bmatrix} \mathrm{Attn}(Q_i^{\mathrm{f}}, K_i^{\mathrm{f}}, V_i^{\mathrm{f}}) + \mathrm{Attn}(Q_i^{\mathrm{f}}, K_t^{\mathrm{f}}, V_t^{\mathrm{f}}) \\ \mathrm{Attn}(Q_t^{\mathrm{f}}, K_i^{\mathrm{f}}, V_i^{\mathrm{f}}) + \mathrm{Attn}(Q_t^{\mathrm{f}}, K_t^{\mathrm{f}}, V_t^{\mathrm{f}}) \end{bmatrix} \tag{4.9}$$

$$\mathrm{Attn}(Q^{\mathrm{m}}, K^{\mathrm{m}}, V^{\mathrm{m}}) = \begin{bmatrix} \mathrm{Attn}(Q_i^{\mathrm{m}}, K_i^{\mathrm{m}}, V_i^{\mathrm{m}}) + \mathrm{Attn}(Q_i^{\mathrm{m}}, K_t^{\mathrm{m}}, V_t^{\mathrm{m}}) \\ \mathrm{Attn}(Q_t^{\mathrm{m}}, K_i^{\mathrm{m}}, V_i^{\mathrm{m}}) + \mathrm{Attn}(Q_t^{\mathrm{m}}, K_t^{\mathrm{m}}, V_t^{\mathrm{m}}) \end{bmatrix} \tag{4.10}$$

This attention mechanism serves as the foundation of our interpolation-based strategy for mitigating bias in Stable Diffusion 3.

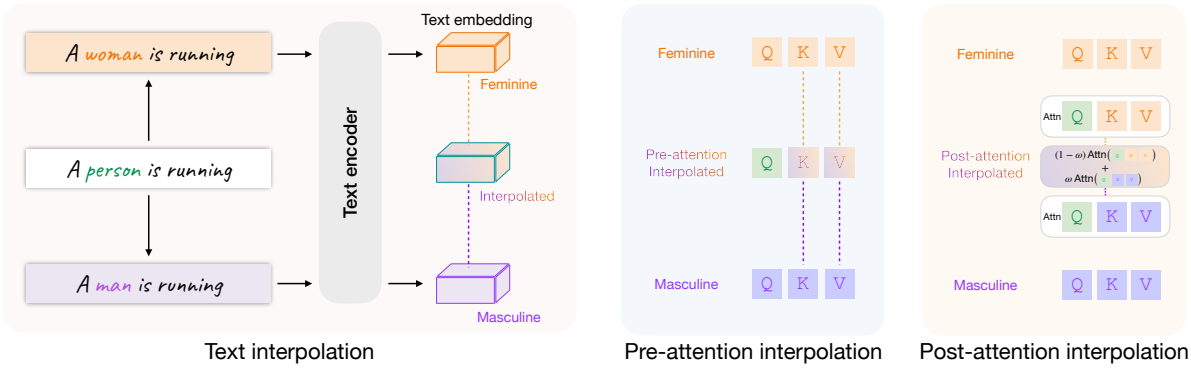

Figure 4.2: Illustration of three interpolation strategies: text interpolation, pre-attention interpolation, and post-attention interpolation.

4.4 Method

Our goal is to mitigate gender bias in text-to-image generation by encouraging the model to produce genuinely neutral images when given neutral prompts. We assume that the concepts of feminine and masculine can span a semantic spectrum, and neutrality can be approximated as a meaningful interpolation between the two. Note that we do not define neutrality by the interpolated images themselves, but rather treat them as representations of neutrality, regardless of the attributes of the generated faces. Given a neutral prompt p_{n}, we construct its feminine and masculine counterparts, p_{f} and p_{m}, respectively. Their corresponding attention computations are given in Eq. 4.8, 4.9, and 4.10.

To mitigate bias in image generation, following [156], we propose three strategies for interpolation that are applied within Stable Diffusion 3 (see Figure 4.2), text embedding interpolation (Sec. 4.4.1), pre-attention interpolation (Sec. 4.4.2), post-attention interpolation (Sec. 4.4.2). The process of parameter selection are further discussed in Sec. 4.4.3.

4.4.1 Text embedding interpolation

As bias originates from the text embedding (as shown in Chapter 3), we first start with a direct manipulation of the input text embedding. The interpolated text embedding c_{r} can be defined

as:

$$c_{\mathrm{r}} = (1-\omega)c_{\mathrm{m}} + \omega c_{\mathrm{f}}, \quad \omega \in [0,1] \tag{4.11}$$

where $\omega \in [0,1]$ is an interpolation weight between the masculine c_{m} and feminine c_{f}.

Since the query, key, and value tensors used in attention computation are linear projections of the text embeddings, the interpolation propagates as:

$$Q_t^{\mathrm{r}} = W_t^{Q_{\mathrm{n}}} c_{\mathrm{r}}, \quad K_t^{\mathrm{r}_t} = W_t^{K_{\mathrm{n}}} c_{\mathrm{r}}, \quad V_t^{\mathrm{r}_t} = W_t^{V_{\mathrm{n}}} c_{\mathrm{r}}, \tag{4.12}$$

where Q_t^{r}, $K_t^{\mathrm{r}_t}$, and $V_t^{\mathrm{r}_t}$ are query, key, and value of the interpolated text embedding.

The attention output is:

$$\mathrm{Attn}_{\text{text\_itp}} = \begin{bmatrix} \mathrm{Attn}(Q_i^{\mathrm{n}}, K_i^{\mathrm{n}}, V_i^{\mathrm{n}}) + \mathrm{Attn}(Q_i^{\mathrm{n}}, K_t^{\mathrm{r}_t}, V_t^{\mathrm{r}_t}) \\ \mathrm{Attn}(Q_t^{\mathrm{r}}, K_i^{\mathrm{n}}, V_i^{\mathrm{n}}) + \mathrm{Attn}(Q_t^{\mathrm{r}}, K_t^{\mathrm{r}_t}, V_t^{\mathrm{r}_t}) \end{bmatrix} \tag{4.13}$$

4.4.2 Attention interpolation

To further refine control over the attention, we introduce two attention-level interpolation methods that operate on top of the text embedding interpolation. These strategies differ in the stage of interpolation, either before or after attention computation.

Pre-attention interpolation

Pre-attention interpolation applies interpolation directly to the key and value before attention is computed. Specifically, we interpolate both visual and textual key-value pairs:

$$K_i^{\mathrm{r}} = (1-\omega)K_i^{\mathrm{m}} + \omega K_i^{\mathrm{f}}, \tag{4.14}$$

$$V_i^{\mathrm{r}} = (1-\omega)V_i^{\mathrm{m}} + \omega V_i^{\mathrm{f}}, \tag{4.15}$$

$$K_t^{\mathrm{r}} = (1-\omega)K_t^{\mathrm{m}} + \omega K_t^{\mathrm{f}}, \tag{4.16}$$

$$V_t^{\mathrm{r}} = (1-\omega)V_t^{\mathrm{m}} + \omega V_t^{\mathrm{f}}. \tag{4.17}$$

The fused attention is then computed using the interpolated key and value tensors:

$$\text{Attn}_{\text{pre}} = \begin{bmatrix} \text{Attn}(Q_i^{\text{n}}, K_i^{\text{r}}, V_i^{\text{r}}) + \text{Attn}(Q_i^{\text{n}}, K_t^{\text{r}}, V_t^{\text{r}}) \\ \text{Attn}(Q_t^{\text{r}}, K_i^{\text{r}}, V_i^{\text{r}}) + \text{Attn}(Q_t^{\text{r}}, K_t^{\text{r}}, V_t^{\text{r}}) \end{bmatrix} \tag{4.18}$$

Post-attention interpolation

In contrast to pre-attention interpolation, post-attention interpolation computes two separate attention outputs using the feminine and masculine key-value pairs, and then interpolates their resulting attention outputs:

$$\text{Attn}_{\text{post}} = (1-\omega) \begin{bmatrix} \text{Attn}(Q_i^{\text{n}}, K_i^{\text{f}}, V_i^{\text{f}}) + \text{Attn}(Q_i^{\text{n}}, K_t^{\text{f}}, V_t^{\text{f}}) \\ \text{Attn}(Q_t^{\text{r}}, K_i^{\text{f}}, V_i^{\text{f}}) + \text{Attn}(Q_t^{\text{r}}, K_t^{\text{f}}, V_t^{\text{f}}) \end{bmatrix} + \omega \begin{bmatrix} \text{Attn}(Q_i^{\text{n}}, K_i^{\text{m}}, V_i^{\text{m}}) + \text{Attn}(Q_i^{\text{n}}, K_t^{\text{m}}, V_t^{\text{m}}) \\ \text{Attn}(Q_t^{\text{r}}, K_i^{\text{m}}, V_i^{\text{m}}) + \text{Attn}(Q_t^{\text{r}}, K_t^{\text{m}}, V_t^{\text{m}}) \end{bmatrix} \tag{4.19}$$

While both pre-attention and post-attention interpolation aim to mitigate bias by balancing semantic influences, they operate at different levels of the attention mechanism and offer different trade-offs. Pre-attention interpolation interpolate the key and value before attention is computed, thereby providing a more entangled and distributed form of control. This allows the attention mechanism to softly align queries with interpolated feature, and it may reduce strong gender-specific signals, potentially leading to overly neutral generation. In contrast, post-attention interpolation computes full attention outputs conditioned on each gender-specific representation independently and interpolates them afterward, preserving the semantic structure of each source more. However, the post interpolation strategy might limit interaction between modalities at a finer level, potentially overlooking the joint nature of MM-DiT. Thus, while post-attention interpolation offers clearer interpretability and control over output composition, pre-attention interpolation may better leverage the capacity of joint attention. The choice between the two depends on the desired trade-off between semantic fidelity and interaction flexibility.

4.4.3 Parameter selection

The interpolation weight ω governs the balance between masculine and feminine semantic representations during image generation. Selecting appropriate values for ω is critical for achieving both semantic neutrality and visual diversity in text-to-image generation. For the all three interpolation strategies introduced above, we apply a unified parameter selection process.

Following [156], we sample the interpolation weights ω from a Beta distribution. This encourages sampling from informative mid-range regions of the interpolation space [0, 1], avoiding overdominance of either gender. Specifically, we define a Beta prior:

$$\omega \sim \text{Beta}(\alpha, \beta), \tag{4.20}$$

where $\alpha, \beta > 1$ yields a bell-shaped distribution. Given a uniform sample $v \sim \mathcal{U}(0, 1)$, we obtain the corresponding interpolation coefficient as:

$$\omega = F_B^{-1}(v; \alpha, \beta), \tag{4.21}$$

where $F_B(\cdot)$ denotes the inverse cumulative distribution function (CDF) of the Beta distribution.

To generate a candidate pool of k interpolation weights $\{\omega_1, \ldots, \omega_k\}$, we first initialize the Beta parameters (α, β) and select a starting interpolation weight ω_{ref} in the middle of the range. The endpoints $\omega = 0$ and $\omega = 1$ correspond to the masculine image I_m and the feminine image I_f, respectively. For each intermediate ω_i, we generate an image I_i and compute its perceptual distances to its neighbors using a CLIP-based distance $P(\cdot, \cdot)$. Let I_{ref} denote the first interpolated image with ω_{ref}. If

$$P(I_{\text{ref}}, I_m) < P(I_{\text{ref}}, I_f), \tag{4.22}$$

we infer that I_{ref} is closer to the masculine and thus sample the next one between I_{ref} to the feminine one, *i.e.*, ω in the interval $(\omega_{\text{ref}}, 1)$. Otherwise, we sample ω from $(0, \omega_{\text{ref}})$. The process applies k times. The Beta parameters α and β are updated iteratively to reflect observed perceptual distances, effectively adapting the prior to the current sampling trajectory.

After generating k interpolated images and computing all pairwise perceptual distances, we select a smooth subsequence of n images $I_{1:n}$ that minimizes the maximum pairwise distance along the path:

$$\min_{1 \le i_1 < \cdots < i_n \le k} \max_{1 \le j < n} \left| P(I_{i_j}, I_{i_{j+1}}) \right|. \tag{4.23}$$

This smooth path selection encourages gradual transitions and filters out samples that cause abrupt semantic shifts, therefore enhancing the stability of neutral image generation.

In practice, we set $k = 1.5n$ to maintain sufficient diversity while keeping the low computational cost. By combining Beta prior sampling with smooth sequence selection, our method efficiently explores the interpolation space and adaptively selects balanced, visually coherent neutral images without the need of exhaustive grid search or model fine-tuning.

4.5 Experiments

In this section, we first present the details of the image generation process (Sec. 4.5.1), then introduce the evaluation metrics for bias mitigation and image quality (Sec. 4.5.2). We then analyze the results in terms of both bias mitigation and generation quality, followed by further ablation analysis.

4.5.1 Image generation details

Prompt preparation We conduct our experiments on four vision-language datasets, GCC validation dataset [140], COCO [137], TextCaps [141], Flickr30k [139], and one sentence set, Profession, which consists of occupational sentences generated by ChatGPT 3.5 [12]. For the vision-language datasets, we extract neutral prompts by selecting captions that contain the word *person* or *people*, while excluding any gender-specific words (*e.g.*, *brother*, *mother*, etc). For the Profession set, which is introduced in Chapter 3, neutral prompts are those that simply mention the profession.

To construct gender-specific counterparts, we modify the neutral prompts. In the vision-language datasets, we substitute *person*/*people* to *woman*/*women* or *man*/*men* to construct feminine and masculine prompts, respectively. For the Profession set, we prepend *female*/*male* on the profession name to form feminine and masculine prompts, respectively. For each dataset, we randomly sample $\mathcal{N} = 100$ neutral prompts and generate images using 3 different seeds per prompt.

Interpolation details Our interpolation method is implemented based on Stable Diffusion 3 medium[2]. Given a neutral prompt, we first generate gendered images using the corresponding feminine and masculine prompts. Then, we create a set of interpolated images by applying interpolation either in the text embedding or within attention mechanism.

We generate $k = 7$ images per prompt, including the interpolation images as well as feminine and masculine images. From these, we select $n = 5$ images that form a perceptually smooth interpolation path. This results in three intermediate interpolations between the feminine and masculine endpoints. All generations share the same initial noise. The inference applies 28 denoising steps.

For beta prior distribution, we set $\alpha = 3$, $\beta = 3$ as initial parameters. The warm-up ratio controls when interpolation is applied during the denoising process. For text interpolation, we apply interpolation across all steps (warm-up ratio is 100%). For attention-based interpolation (pre-attention and post-attention), we set the warm-up ratio to 20%, maintaining high-level semantic information and high-quality of the generations.

4.5.2 Metrics

We evaluate our method using two categories of metrics: gender bias mitigation evaluation and image quality evaluation. The goal is to quantify both the effectiveness of our interpolation strategies in mitigating bias and their impact on visual quality and coherence.

[2] `https://huggingface.co/stabilityai/stable-diffusion-3-medium`

Bias evaluation

To evaluate gender bias, we compute cosine similarity across the three spaces during the generation process: prompt space, denoising space, and image space, following Chapter 3. Based on these similarities, we compute a bias mitigation score that reflects how effectively bias is mitigated.

Prompt space We define the prompt space using the text embeddings of the input prompts. Specifically, we compute cosine similarity between either the original neutral embedding c_{n} or the interpolated embedding c_{r} and the gendered embeddings c_{f} (feminine) and c_{m} (masculine). The average similarity across all prompts is:

$$s_{\mathrm{P}}(\text{neu/interpolation}, \text{fem/mas}) = \frac{1}{|\mathcal{N}|} \sum \cos(c, c'), \quad c \in \{c_{\mathrm{n}}, c_{\mathrm{r}}\}, c' \in \{c_{\mathrm{f}}, c_{\mathrm{m}}\} \tag{4.24}$$

where $\mathcal{N}$ is the number of neutral prompts.

Denoising space The denosing space is taken from the last step of the denoising process when $T = 0$. The cosine similarity in denoising space can be presented as:

The denoising space corresponds to the latent embeddings at the final denoising step ($T = 0$). We compute cosine similarity between the embedding of neutral or interpolated images and those of the gendered images:

$$s_{\mathrm{D}}(\text{neu/interpolation}, \text{fem/mas}) = \frac{1}{|\mathcal{N}|} \sum \cos(\mathbf{z}_0, \mathbf{z}_0'), \quad \mathbf{z}_0 \in \{\mathbf{z}_0^{\mathrm{n}}, \mathbf{z}_0^{\mathrm{r}}\}, \mathbf{z}_0' \in \{\mathbf{z}_0^{\mathrm{f}}, \mathbf{z}_0^{\mathrm{m}}\}, \tag{4.25}$$

where $\mathbf{z}_0^{\mathrm{n}}$, $\mathbf{z}_0^{\mathrm{r}}$, $\mathbf{z}_0^{\mathrm{f}}$, and $\mathbf{z}_0^{\mathrm{m}}$ are the embeddings in the denoising space for the neutral, interpolated, feminine, and masculine images, respectively.

Image space To capture perceptual and semantic differences at the image level, we compute cosine similarity in the image space using multiple complementary methods. SSIM evaluates

structural similarity, measuring pixel-level correspondence. For high-level semantic similarity, we extract latent features using pre-trained models, the final layer of ResNet-50 [71], the image encoder from CLIP ViT-B/32[3] [26], and the final layer of DINO-s16 [52]. These are referred to as ResNet, CLIP, and DINO, respectively. We then compute the cosine similarity between the interpolated (or neutral) image and each gendered counterpart to quantify bias at different representational levels.

Bias mitigation score We define a bias mitigation score $\mathcal{B}$ to quantify the extent to which bias is mitigated through interpolation. Let x represents embeddings either the neutral embedding $\mathcal{X}_{\text{neu}}$ or the interpolated embedding $\mathcal{X}_{\text{itp}}$, where $\mathcal{X}$ denotes the latent embeddings of all the samples in the given space. The bias distance $\mathcal{D}_{\text{space}}(x)$ is computed as:

$$\mathcal{D}_{\text{space}}(x) = \left| s_{\text{space}}(x, \mathcal{X}_{\text{fem}}) - s_{\text{space}}(x, \mathcal{X}_{\text{mas}}) \right|, \quad x \in \{\mathcal{X}_{\text{neu}}, \mathcal{X}_{\text{itp}}\}, \tag{4.26}$$

where $\mathcal{X}_{\text{fem}}$ and $\mathcal{X}_{\text{mas}}$ denote the embedding of the feminine and masculine, respectively.

The bias mitigation score is then defined as the reduction in bias distance after applying interpolation:

$$\mathcal{B} = \mathcal{D}_{\text{space}}(\mathcal{X}_{\text{neu}}) - \mathcal{D}_{\text{space}}(\mathcal{X}_{\text{itp}}). \tag{4.27}$$

A positive value of $\mathcal{B}$ indicates that the interpolated image exhibits more balanced similarity to both gendered counterparts, thereby mitigating bias and yielding a more neutral representation. Conversely, a negative $\mathcal{B}$ suggests that interpolation has amplified bias.

Generation evaluation

To evaluate the quality and behavior of generated interpolations, we adopt three metrics from [156]: Consistency, Smoothness, and Fidelity. These metrics evaluate whether the generated interpolations produce high-quality, coherent sequences while maintaining high fidelity to the original generation distribution.

[3] https://github.com/openai/CLIP

Consistency Consistency measures local coherence between adjacent interpolated images. Let $P(\cdot, \cdot)$ denote perceptual distance measured using LPIPS [182]. We use CLIP as the extractor. For a sequence of n interpolated images $I_{1:n}$, the consistency score $\mathcal{C}$ is defined as:

$$\mathcal{C}\left(I_{1:n}; P\right) = \frac{1}{n-1} \sum_{i=1}^{n-1} P\left(I_i, I_{i+1}\right). \tag{4.28}$$

A lower value of $\mathcal{C}$ indicates that each image is perceptually close to its neighbors, suggesting a smooth and semantically coherent transition. This ensures that the selected neutral image arises from a trajectory that follows the visual semantics of both gendered images.

Smoothness While consistency captures local similarity between adjacent images, the smoothness score $\mathcal{S}$ evaluates the uniformity of change across the entire interpolation sequence. It is defined using the Gini coefficient $\mathcal{G}(D)$ computed over the LPIPS distances between neighboring image pairs. Let the LPIPS distances be denoted as:

$$D = \{d_1, \cdots, d_{n-1}\}, \text{where } d_i = P(I_i, I_{i+1}). \tag{4.29}$$

Then the Gini coefficient $\mathcal{G}(D)$ is given by:

$$\mathcal{G}(D) = \frac{\sum_{i=1}^{n-1} \sum_{j=1}^{n-1} |d_i - d_j|}{2(n-1) \sum_{i=1}^{n-1} d_i}, \tag{4.30}$$

and the smoothness score $\mathcal{S}$ is defined as:

$$\mathcal{S}(I_{1:n}; P) = 1 - \mathcal{G}(D). \tag{4.31}$$

A higher smoothness score $\mathcal{S}$ indicates more uniformly perceptual changes throughout the sequence, suggesting that interpolation progresses at a steady semantic pace without abrupt transitions.

Fidelity Fidelity evaluates how closely interpolated images match the distribution of the images generated from the original Stable Diffusion 3. We use the Fréchet Inception Distance (FID) [133] between the set of interpolated images and the original gendered endpoints. Let

$\mathcal{I}_{\text{end}} = \{I_{\text{f}}, I_{\text{m}}\}$ and $\mathcal{I}_{\text{itp}}$ be the interpolated set (excluding endpoints). The FID score $\mathcal{F}$ is defined as:

$$\mathcal{F} = \text{FID}(\mathcal{I}_{\text{end}}, \mathcal{I}_{\text{itp}}). \tag{4.32}$$

A lower FID indicates that the interpolated images are closer in distribution to those generated from the feminine and masculine prompts, implying better visual fidelity.

4.5.3 Results analysis

Bias in Stable Diffusion 3

We begin by evaluating gender bias in the original Stable Diffusion 3 by computing the similarity between neutral and gendered images across the three spaces. Results in Table 4.1 show that the difference in the prompt space is nearly zero, indicating that the neutral and feminine/masculine text embeddings are similar. However, noticeable disparities remain in both denoising and image spaces, where neutral representations are consistently closer to masculine ones. This reflects that gender bias persists in Stable Diffusion 3, aligning with observations in Chapter 3, and motivates the need for bias mitigation.

Bias mitigation evaluation

We evaluate the effectiveness of our bias mitigation strategies using the bias mitigation score $\mathcal{B}$, computed across three spaces (prompt, denoising, image) and five datasets, under three interpolation methods: text, pre-attention, and post-attention interpolation.

Figure 4.3 shows results where the neutral image is randomly selected from the interpolation images. All three interpolation methods consistently mitigate bias in the prompt space across all datasets. In most cases, bias is also mitigated in the denoising and image spaces, with the exception of the CLIP space and the Profession dataset.

Surprisingly, the Profession dataset exhibits the lowest bias initially in most spaces (except CLIP), suggesting that prompts containing professions may be inherently more balanced.

dataset	pairs	text space	denoising space	image space			
				SSIM	ResNet	CLIP	DINO
GCC	(neu, fem)	0.9984	0.7410	0.5120	0.8369	0.7739	0.6662
	(neu, mas)	0.9984	**0.7545**	**0.5223**	**0.8445**	**0.7747**	**0.6764**
COCO	(neu, fem)	0.9985	0.7582	0.5139	0.8168	**0.8148**	0.7221
	(neu, mas)	**0.9986**	**0.7725**	0**.5278**	**0.8725**	0.8064	**0.7296**
TextCaps	(neu, fem)	**0.9988**	**0.7282**	0.4887	0.8489	0.7589	0.6948
	(neu, mas)	0.9987	0.7528	**0.4986**	**0.8621**	**0.7672**	**0.7178**
Flickr30k	(neu, fem)	0.9984	0.7835	0.5005	0.8508	**0.8058**	0.7130
	(neu, mas)	**0.9986**	**0.8003**	**0.5227**	**0.8730**	0.8020	**0.7390**
Profession	(neu, fem)	0.9933	0.7169	**0.5100**	0.8242	**0.8243**	0.6202
	(neu, mas)	0.9933	**0.7251**	0.5063	**0.8276**	0.8221	**0.6242**

Table 4.1: Cosine similarity between the neutral prompts and the counterpart prompts (feminine and masculine) across different spaces including text, denoising, and image (SSIM, ResNet, CLIP, and DINO). Results are computed using Stable Diffusion 3.

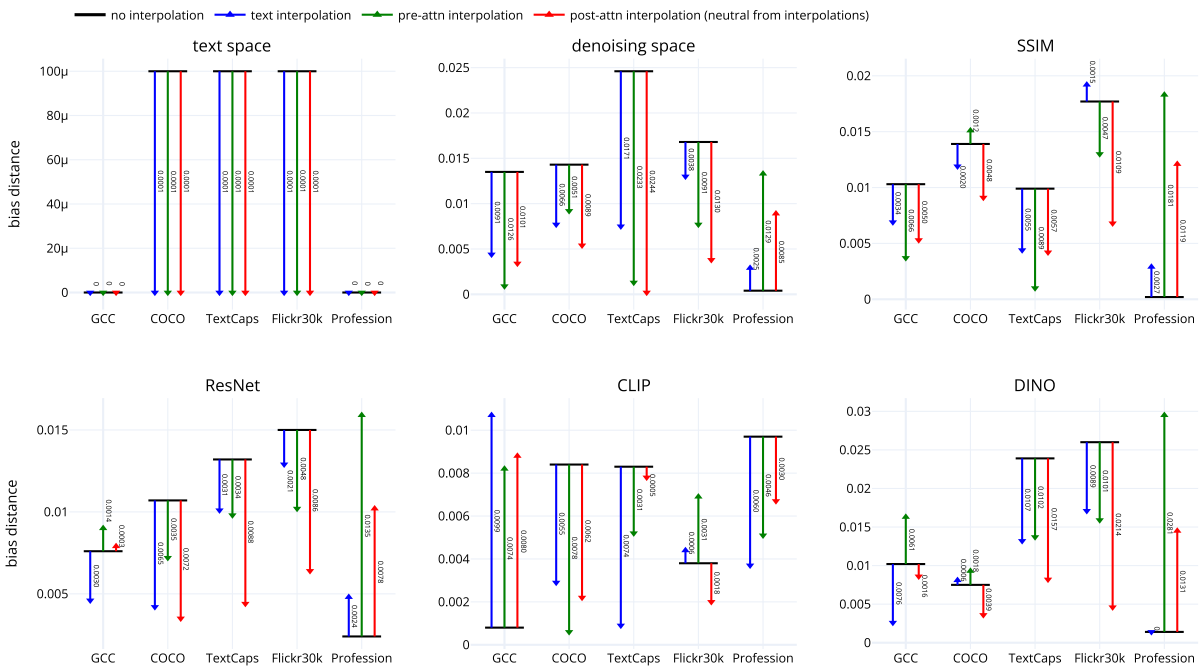

Figure 4.3: Bias distance across three interpolation strategies when neutral images are selected randomly from the interpolations. The black line represents the original bias distance $\mathcal{D}_{\text{space}}(\mathcal{X}_{\text{neu}})$ in Stable Diffusion 3, while arrow endpoints indicate the bias distance $\mathcal{D}_{\text{space}}(\mathcal{X}_{\text{itp}})$ after applying mitigation. A downward arrow corresponds to a positive bias mitigation score $\mathcal{B}$, indicating bias is successfully mitigated.

However, after interpolation, the bias distance occasionally increases. This may be due to the construction of gendered prompts in this set – by prepending *female* or *male* – which may not yield effective counterparts for this interpolation-based mitigation.

To explore a broader candidate pool for neutrality, we additionally include masculine and feminine images alongside interpolations when selecting the neutral output. As shown in Figure 4.4, this setup may sometimes increase bias, which may be caused when gendered images are selected as neutral. The results highlight the importance of using interpolated representations that are explicitly designed to balance between the gendered endpoints, rather than relying on original gendered images.

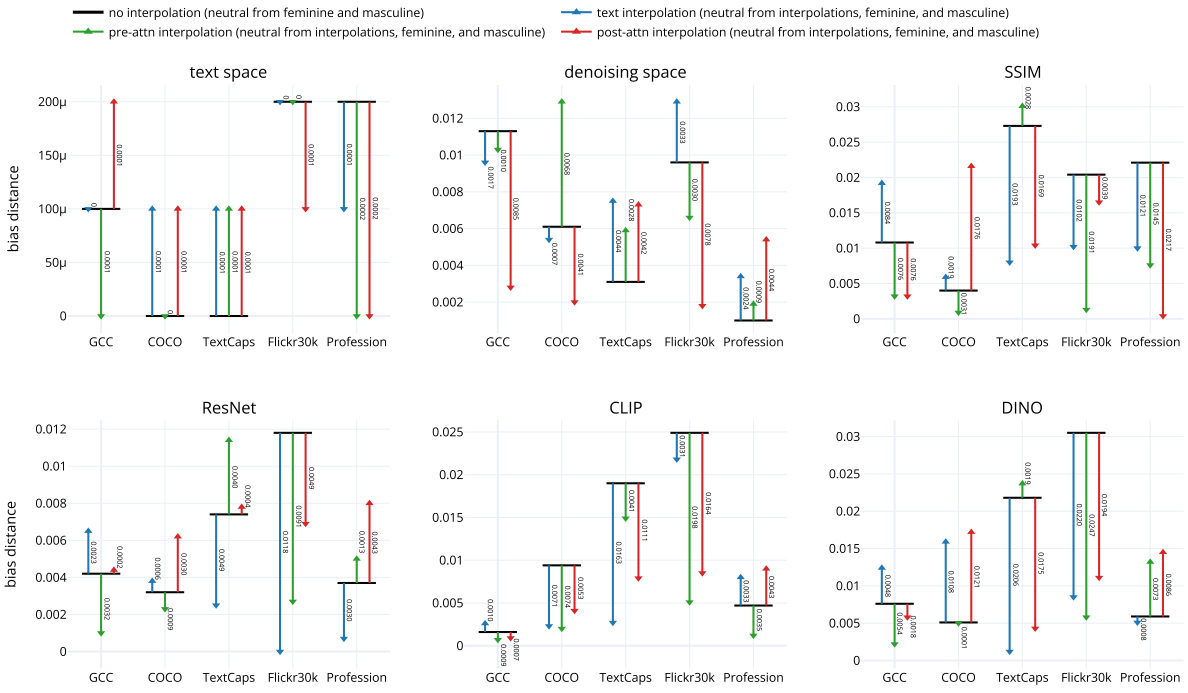

Figure 4.4: Bias distance when neutral images are sampled from interpolations, as well as from original feminine and masculine images. A downward arrow indicates mitigated bias compared to the original Stable Diffusion 3 (black line).

Image generation evaluation

To evaluate the visual quality of the interpolated neutral images, we compute Consistency, Smoothness, and Fidelity across all datasets. The results are presented in Table 4.2.

Across the five datasets, a consistent trend emerges: text interpolation achieves the highest smoothness, while post-attention interpolation performs best in terms of both consistency and fidelity. Text interpolation operates solely on the text embedding, which results in perceptually uniform transitions without abrupt semantic shifts, thereby yielding smoother interpolations. In contrast, post-attention interpolation modifies the output of the attention mechanism, preserving structural coherence and generating outputs that are more similar to those of the original Stable Diffusion 3 model.

Pre-attention interpolation performs comparably to post-attention in both consistency and

smoothness, suggesting that interpolating key and value is effective. However, post-attention interpolation slightly outperforms it in consistency and fidelity, indicating that manipulating the attention output directly may more effectively maintain structural integrity and overall visual quality.

Qualitative results

Figure 4.5 presents the neutral images generated using our bias mitigation method across three interpolation strategies. The results demonstrate that our method successfully produces individuals with diverse demographic characteristics. Notably, it generates uncommon combinations such as a person with masculine facial features wearing a dress (middle image in the third and fourth rows on the left), highlighting a broader and more inclusive range of attributes. In the example on the right, which involves two people, our method maintains high image quality while also generating underrepresented demographics such as elderly individuals, an aspect typically absent in neutral prompt generations.

Warm-up ratio

In diffusion models, earlier denoising steps tend to encode high-level semantic information, while later steps focus on refining low-level details. To balance image quality and bias mitigation, we introduce a warm-up ratio that controls how many denoising steps interpolation is applied to.

Figure 4.6 illustrates the effect of different warm-up ratios in post-attention interpolation. As the number of interpolated steps increases, particularly above 50%, visual artifacts emerge and become increasingly apparent. This suggests that excessive interpolation disrupts the model's ability to preserve image fidelity.

To identify an optimal trade-off between image quality and bias mitigation, we evaluate the bias distance on post-attention interpolation under three warm-up ratios: $10\%, 20\%$, and 30%. As shown in Figure 4.7, the 20% setting consistently achieves the lowest bias across all evaluated spaces. Based on these results, we adopt 20% as the default warm-up ratio in all main

dataset	interpolation	consistency ↓	smoothness ↑	fidelity ↓
GCC	text	0.2431	**0.8387**	43.030
	pre-attn	0.2078	0.8178	41.260
	post-attn	**0.2044**	0.8188	**41.180**
COCO	text	0.2140	**0.8425**	43.920
	pre-attn	**0.1834**	0.8124	43.330
	post-attn	0.1836	0.8263	**43.051**
TextCaps	text	0.2494	**0.8440**	46.187
	pre-attn	0.2069	0.8265	**43.866**
	post-attn	**0.2024**	0.8207	43.870
Flickr30k	text	0.2103	**0.8363**	45.316
	pre-attn	0.1822	0.8169	43.863
	post-attn	**0.1813**	0.8225	**43.476**
Profession	text	0.2380	**0.8478**	42.055
	pre-attn	0.1990	0.8149	41.473
	post-attn	**0.1982**	0.8214	**41.313**

Table 4.2: Evaluation of interpolation quality across datasets. Lower consistency indicates better visual coherence between neighboring images, higher smoothness reflects more uniform transitions across the interpolations, and lower FID suggests higher fidelity to the original data distribution.

experiments.

4.6 Limitations

While our method offers a lightweight and training-free approach to generate diverse neutral images for mitigating gender bias in text-to-image generation, several limitations remain.

Limited visual attribute control Our method does not control every element in the generated image. Although we interpolate between gendered images to increase neutrality at the image level, the method does not support fine-grained control over specific visual attributes such as clothing (*e.g.*, suits or dresses) or background context (*e.g.*, presence of an audience). The objective is to make the overall image appear neither skewed toward feminine nor masculine, rather than focusing only on limited elements like facial features. As such, some unintended associations may still persist in the generated image.

Lack of demographic attribute assignment While we aim for neutrality by interpolating across gendered representations, we do not have control over other demographic characteristics (*e.g.*, race, or age) of the faces generated in the images. However, this is also a fundamental limitation of the original neutral prompts themselves, which are inherently not tailored to any specific attributes.

Indirect neutral image generation Our method generates multiple candidate images and selects one to promote diversity and reduce bias. However, it does not directly predict the distribution of neutral images. Learning the distribution of neutral to improve efficiency is one of the directions for future work.

Discontinuity in interpolation As the distribution of representations may be not continuous, applying interpolation between two points (particularly in the text space) may result in abrupt

changes in the output. Although attention interpolation helps smooth the transitions, this issue still persists and requires further exploration in the future.

Limited to gender bias As our method do not fine-tuning the model and targets only gender bias, the inherent bias of the pre-trained model persist and other bias including race, age, and cultural bias may not be fully mitigated through our interpolation strategy alone. Extending this strategy to adopt to a wider range of societal bias is an important direction for future research.

Limited evaluation metrics The evaluation of bias mitigation in generative models remains inherently challenging. While we adopt quantitative metrics such as LPIPS and bias distance, these do not fully reflect the bias or diversity of all the elements in the image, such as the outfits, the background, etc. A more holistic evaluation framework would be beneficial.

Dependency on attention module Our method is designed for models that contain attention mechanisms, particularly the MM-DiT backbone used in Stable Diffusion 3. As the interpolation is applied to the attention module, it may not be directly applicable to architectures that do not reply on attention (*e.g.*, GAN-based or diffusion models with other conditioning strategies).

4.7 Summary

In this chapter, we proposed a training-free method to mitigate gender bias in text-to-image generation. By interpolating between feminine and masculine embeddings within both the text embedding and attention modules of Stable Diffusion 3, our approach generates fairer and more diverse outputs for neutral prompts. This method does not require model fine-tuning or additional data, making it lightweight and easy to integrate into existing Stable Diffusion 3-based models. Among the three proposed interpolation strategies explored, post-attention interpolation yields the best performance. Our experiments demonstrated that our approach effectively reduces representational disparities while maintaining high image quality and semantic alignment.

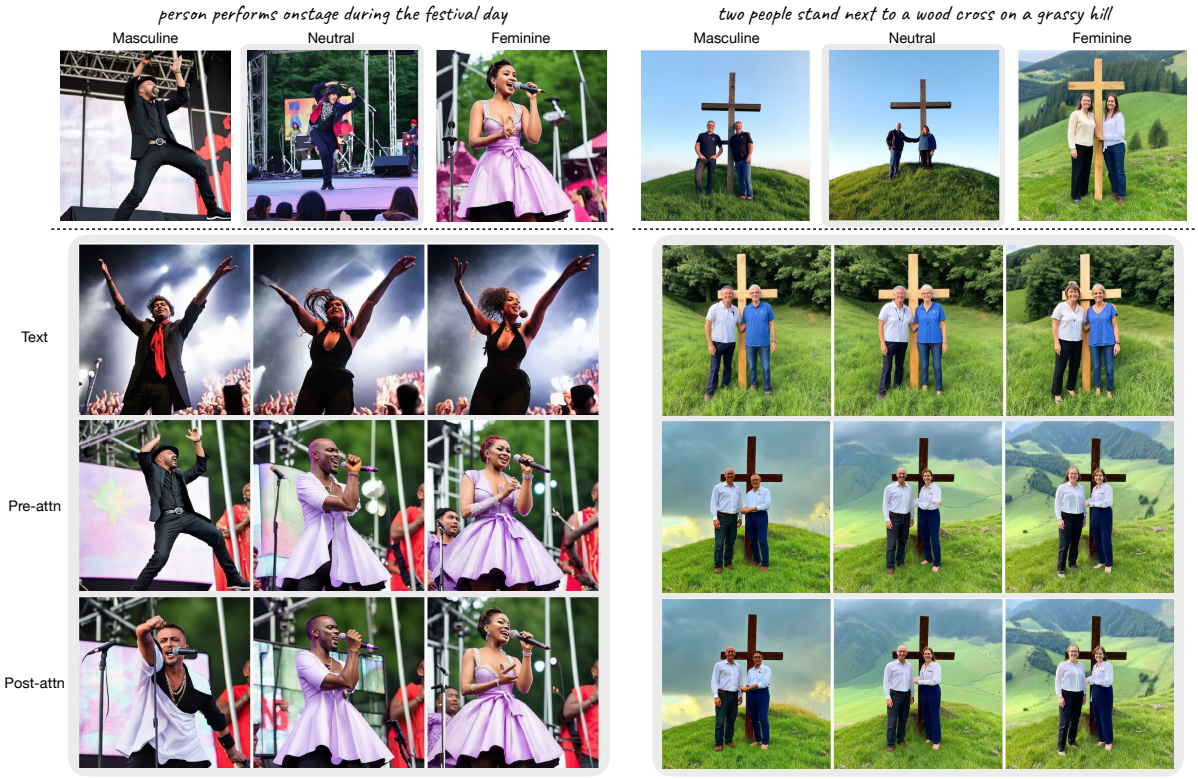

Figure 4.5: Qualitative results on the neutral images generated from three strategies: text interpolation (second row), pre-attention interpolation (third row), and post-attention interpolation (fourth row). The top row shows images generated by the original Stable Diffusion 3 using the masculine, neutral, and feminine prompts. The examples illustrate that our method enables the synthesis of demographically diverse and high-quality outputs, including uncommon combinations of attributes and underrepresented groups. The neutral prompts are shown above the examples.

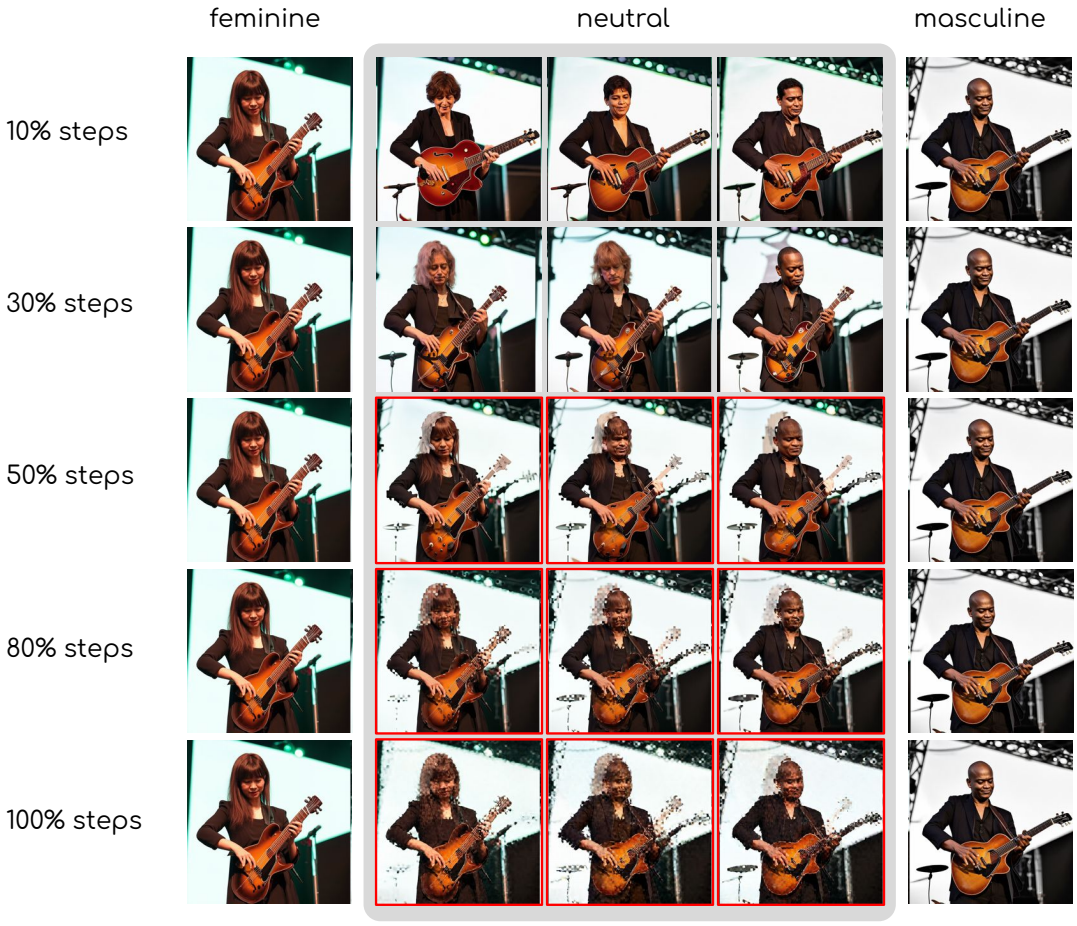

Figure 4.6: Interpolations under different warm-up ratios ($10\%, 30\%, 50\%,$ and 80%) in post-attention interpolation. More steps are interpolated (especially over 50%), more artifacts appear in the output images.

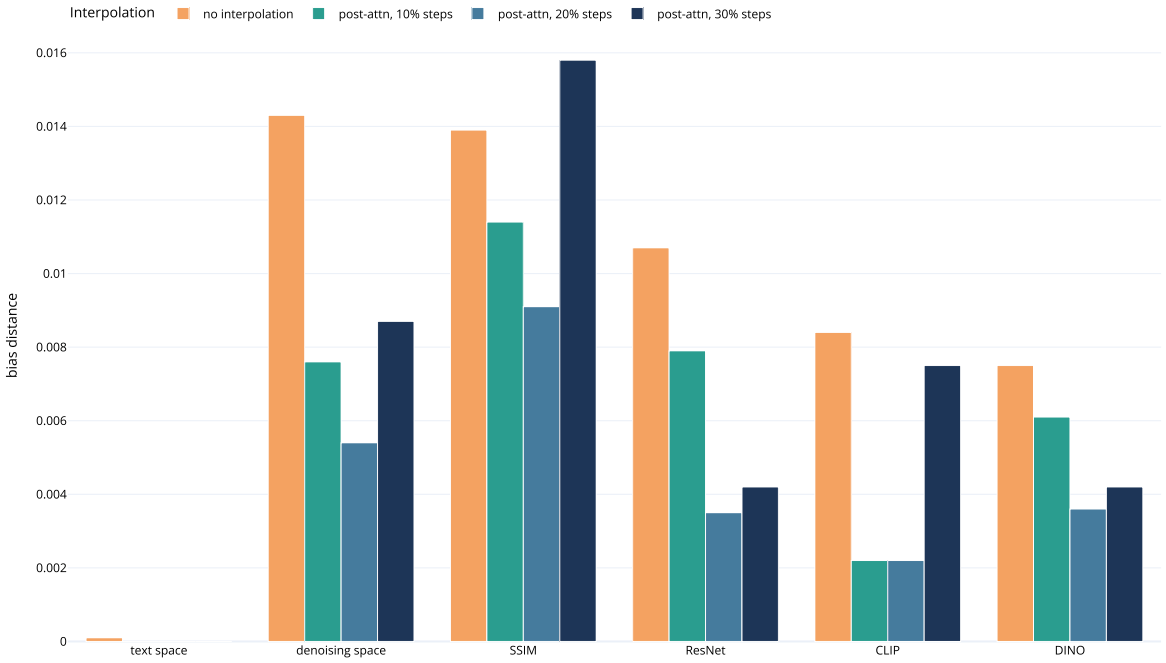

Figure 4.7: Bias distance across three warm-up ratios $(10\%, 20\%, 30\%)$ in post-attention interpolation, evaluated on the COCO dataset. Lower values indicate better bias mitigation. The baseline (no interpolation) is shown in orange.

Chapter 5

Conclusion

This thesis studied the impact of text-to-image models on art and society. On the one hand, we explored generative models as a powerful tool for artistic understanding. On the other hand, we introduced a critical evaluation of gender bias in generative models and further proposed a lightweight approach to mitigate bias without model fine-tuning. Investigating both artistic analysis and social responsibility, this work offered a comprehensive investigation into how generative models influence art and society.

To explore the use of generative models for art analysis, we proposed GOYA, a novel framework that leverages synthetic data from Stable Diffusion to disentangle content and style with contrastive learning. Our evaluation demonstrates the effectiveness of the method for downstream tasks such as classification and retrieval, highlighting the potential of generative models in contributing to the digital humanities.

To address the ethical considerations, we introduced an automatic protocol for evaluating gender bias in text-to-image generation. We examine representational disparity across several stages of generation, object co-occurrence in the images, and prompt-image dependencies. Our results show that bias originates from text embedding, perpetuates throughout the generation process, and manifests across the entire image.

Building on these findings, we further presented a training-free interpolation method to mitigate gender bias by interpolating feminine and masculine embeddings within both the text

embedding and attention module of Stable Diffusion 3. This plug-and-play approach reduces disparities while preserving image quality, offering a practical solution for fairer and diverse outputs.

In conclusion, these contributions highlight the potential of generative models not only for creating visual content but also for representation learning. However, as these models contain inherent bias, we should be cautious when using them. We hope that the findings and methods in the paper serve not only as a foundation for future research but also as a guide for the ethical development and application of text-to-image models in society.

Acknowledgements

Applying to the PhD program at this lab and working under the supervision of my dearest supervisors has been one of the best decisions of my life.

At first, I would like to express my deepest gratitude to my supervisors, Associate Professor Noa Garcia and Professor Yuta Nakashima. I am truly proud to be their student. Their constant support, encouragement, and insightful feedback have guided me throughout my doctoral journey. They have been the lighthouse in my academic life, lighting my way whenever I found myself in darkness.

Thanks to Associate Professor Noa Garcia, I was fortunate to have the invaluable opportunity to undertake an internship under the supervision of Associate Professor Giorgos Tolias at Czech Technical University in Prague. I also deeply appreciate Associate Professor Giorgos Tolias for his kind support and generous hospitality during my stay. This experience was both academically enriching and personally memorable.

Moreover, I would like to thank my collaborators for their support and dedication throughout our joint research. In particular, I like to express my appreciation to Dr. Giorgos Kordopatis-Zilos, Dr. Zakaria Laskar, Dr. Yusuke Hirota, Assistant Professor Amelia Katirai, and Ms. Rawisara Lohanimit. Collaborating with them has been both productive and inspiring, and their contributions have been a vital part of this journey.

My sincere thanks also go to the faculty members of our lab, including Professor Hajime Nagahara, Professor Yuta Nakashima, Associate Professor Noa Garcia, and Associate Professor Hideaki Hayashi. Their advice and encouragement have greatly deepened my research and academic development.

I also want to express my heartfelt thanks to my undergraduate supervisors, Professor Ping Zhong and Professor Yanfei Wang. I would not have begun my research journey without their early support and consistent encouragement, and the valuable experience of working with them. Their mentorship inspired me to pursue academic research.

I am deeply grateful to the committee of the fellowship office at The University of Osaka for the funding and financial support of the 次世代挑戦的研究者育成プロジェクト. Their funding made it possible for me to devote myself fully to this research.

Lastly, I would like to thank my family and friends. Their love, patience, and support gave me the strength to overcome difficulties and the courage to continue moving forward.

Bibliography

[1] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In ICML, 2024.

[2] Wei Ren Tan, Chee Seng Chan, Hernan Aguirre, and Kiyoshi Tanaka. Improved ArtGAN for conditional synthesis of natural image and artwork. Transactions on Image Processing, 2019.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In CVPR, 2022.

[4] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125, 2022.

[5] Xingchao Peng, Baochen Sun, Karim Ali, and Kate Saenko. Learning deep object detectors from 3d models. In ICCV, 2015.

[6] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In CVPR, 2019.

[7] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. StableRep: Synthetic images from text-to-image models make strong visual representation learners. NeurIPS, 2024.

[8] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. In NeurIPS Datasets and Benchmarks Track, 2023.

[9] Eslam Mohamed Bakr, Pengzhan Sun, Xiaogian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. HRS-Bench: Holistic, reliable and scalable benchmark for text-to-image models. In ICCV, 2023.

[10] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In FAccT, 2023.

[11] Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. Contrastive language-vision ai models pretrained on web-scraped multimodal data exhibit sexual objectification bias. In FAccT, 2023.

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In NeurIPS, 2020.

[13] Gustavo Carneiro, Nuno Pinho da Silva, Alessio Del Bue, and João Paulo Costeira. Artistic image classification: An analysis on the printart database. In ECCV, 2012.

[14] Noa Garcia, Benjamin Renoust, and Yuta Nakashima. Context-aware embeddings for automatic art analysis. In ICMR, 2019.

[15] Eva Cetinic, Tomislav Lipic, and Sonja Grgic. Fine-tuning convolutional neural networks for fine art classification. Expert Systems with Applications, 2018.

[16] Nanne Van Noord, Ella Hendriks, and Eric Postma. Toward discovery of the artist's style: Learning to recognize artists by their artworks. IEEE Signal Processing Magazine, 2015.

[17] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In CVPR, 2016.

[18] Nikolaos-Antonios Ypsilantis, Noa Garcia, Guangxing Han, Sarah Ibrahimi, Nanne Van Noord, and Giorgos Tolias. The Met dataset: Instance-level recognition for artworks. In NeurIPS Datasets and Benchmarks Track, 2021.

[19] Sabine Lang and Bjorn Ommer. Reflecting on how artworks are processed and analyzed by computer vision. In ECCV Workshops, 2018.

[20] Noa Garcia, Benjamin Renoust, and Yuta Nakashima. ContextNet: representation and exploration for painting classification and retrieval in context. International Journal of Multimedia Information Retrieval, 2020.

[21] Tianwei Chen, Noa Garcia, Liangzhi Li, and Yuta Nakashima. Retrieving emotional stimuli in artworks. In ICMR, 2024.

[22] Zechen Bai, Yuta Nakashima, and Noa Garcia. Explain me the painting: Multi-topic knowledgeable art description generation. In ICCV, 2021.

[23] Yankun Wu, Yuta Nakashima, and Noa Garcia. Not only generative art: Stable diffusion for content-style disentanglement in art analysis. In ICMR, 2023.

[24] Yankun Wu, Yuta Nakashima, and Noa Garcia. Goya: Leveraging generative art for content-style disentanglement. Journal of Imaging, 2024.

[25] Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In CVPR, 2023.

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In ICML, 2021.

[27] Wei Ren Tan, Chee Seng Chan, Hernán E Aguirre, and Kiyoshi Tanaka. Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification. In ICIP, 2016.

[28] Cheikh Brahim El Vaigh, Noa Garcia, Benjamin Renoust, Chenhui Chu, Yuta Nakashima, and Hajime Nagahara. GCNBoost: Artwork classification by label propagation through a knowledge graph. In ICMR, 2021.

[29] Nicolas Gonthier, Yann Gousseau, Said Ladjal, and Olivier Bonfait. Weakly supervised object detection in artworks. In ECCV Workshops, 2018.

[30] Xi Shen, Alexei A Efros, and Mathieu Aubry. Discovering visual patterns in art collections with spatially-consistent feature learning. In CVPR, 2019.

[31] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. International Journal for Digital Art History, 2016.

[32] Hui Mao, Ming Cheung, and James She. DeepArt: Learning joint representations of visual arts. In ACM MM, 2017.

[33] Thomas Mensink and Jan Van Gemert. The rijksmuseum challenge: Museum-centered visual recognition. In ICMR, 2014.

[34] Michael J Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge Belongie. BAM! the behance artistic media dataset for recognition beyond photography. In ICCV, 2017.

[35] Gjorgji Strezoski and Marcel Worring. OmniArt: a large-scale artistic benchmark. TOMM, 2018.

[36] Selina J. Khan and Nanne van Noord. Stylistic multi-task analysis of ukiyo-e woodblock prints. In BMVC, 2021.

[37] Wei-Ta Chu and Yi-Ling Wu. Image style classification based on learnt deep correlation features. Transactions on Multimedia, 2018.

[38] Matthia Sabatelli, Mike Kestemont, Walter Daelemans, and Pierre Geurts. Deep transfer learning for art classification problems. In ECCV Workshops, 2018.

[39] Catherine Sandoval, Elena Pirogova, and Margaret Lech. Two-stage deep learning approach to the classification of fine-art paintings. IEEE Access, 2019.

[40] Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Bjorn Ommer. Content and style disentanglement for artistic style transfer. In ICCV, 2019.

[41] Xin Xie, Yi Li, Huaibo Huang, Haiyan Fu, Wanwan Wang, and Yanqing Guo. Artistic style discovery with independent components. In CVPR, 2022.

[42] Yichun Shi, Xiao Yang, Yangyue Wan, and Xiaohui Shen. SemanticStyleGAN: Learning compositional generative priors for controllable image synthesis and editing. In CVPR, 2022.

[43] Zipeng Xu, Tianwei Lin, Hao Tang, Fu Li, Dongliang He, Nicu Sebe, Radu Timofte, Luc Van Gool, and Errui Ding. Predict, prevent, and evaluate: Disentangled text-driven image manipulation empowered by pre-trained vision-language model. In CVPR, 2022.

[44] Xiaoming Yu, Yuanqi Chen, Shan Liu, Thomas Li, and Ge Li. Multi-mapping image-to-image translation via learning disentanglement. NeurIPS, 2019.

[45] Aviv Gabbay and Yedid Hoshen. Improving style-content disentanglement in image-to-image translation. arXiv preprint arXiv:2007.04964, 2020.

[46] Emily L Denton, et al. Unsupervised learning of disentangled representations from video. NeurIPS, 2017.

[47] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In ICLR, 2017.

[48] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. ICLR, 2023.

[49] Dan Ruta, Saeid Motiian, Baldo Faieta, Zhe Lin, Hailin Jin, Alex Filipkowski, Andrew Gilbert, and John Collomosse. ALADIN: all layer adaptive instance normalization for fine-grained style similarity. In ICCV, 2021.

[50] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In ICCV, 2017.

[51] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing ViT features for semantic appearance transfer. In CVPR, 2022.

[52] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In ICCV, 2021.

[53] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. CogView: Mastering text-to-image generation via transformers. In NeurIPS, 2021.

[54] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In CVPR, 2022.

[55] Zhiheng Li, Martin Renqiang Min, Kai Li, and Chenliang Xu. StyleT2I: Toward compositional and high-fidelity text-to-image synthesis. In CVPR, 2022.

[56] Gihyun Kwon and Jong Chul Ye. CLIPstyler: Image style transfer with a single text condition. In CVPR, 2022.

[57] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. DF-GAN: A simple and effective baseline for text-to-image synthesis. In CVPR, 2022.

[58] Hongchen Tan, Xiuping Liu, Meng Liu, Baocai Yin, and Xin Li. KT-GAN: knowledge-transfer generative adversarial network for text-to-image synthesis. Transactions on Image Processing, 2020.

[59] Amelia Katirai, Noa Garcia, Kazuki Ide, Yuta Nakashima, and Atsuo Kishimoto. Situating the social issues of image generation models in the model life cycle: a sociotechnical approach. arXiv preprint arXiv:2311.18345, 2023.

[60] Johann Ostmeyer, Ludovica Schaerf, Pavel Buividovich, Tessa Charles, Eric Postma, and Carina Popovici. Synthetic images aid the recognition of human-made art forgeries. Plos one, 2024.

[61] Ryuichiro Hataya, Han Bao, and Hiromi Arai. Will large-scale generative models corrupt future datasets? In ICCV, 2023.

[62] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. arXiv preprint arXiv:2304.08466, 2023.

[63] Tianwei Chen, Yusuke Hirota, Mayu Otani, Noa Garcia, and Yuta Nakashima. Would deep generative models amplify bias in future models? In CVPR, 2024.

[64] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. NeurIPS, 2019.

[65] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In NeurIPS, 2020.

[66] Ruizhe Cheng, Bichen Wu, Peizhao Zhang, Peter Vajda, and Joseph E Gonzalez. Data-efficient language-supervised zero-shot learning with self-distillation. In CVPR, 2021.

[67] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. PointCLIP: Point cloud understanding by CLIP. In CVPR, 2022.

[68] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. DiffusionCLIP: Text-guided diffusion models for robust image manipulation. In CVPR, 2022.

[69] LA Gatys, AS Ecker, and M Bethge. A neural algorithm of artistic style. Nature Communications, 2015.

[70] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In ICML, 2020.

[71] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.

[72] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. NeurIPS, 2022.

[73] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. ICLR, 2022.

[74] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In ICLR, 2015.

[75] Xiao Liu, Spyridon Thermos, Gabriele Valvano, Agisilaos Chartsias, Alison O'Neil, and Sotirios A Tsaftaris. Measuring the biases and effectiveness of content-style disentanglement. BMVC, 2021.

[76] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In CVPR, 2021.

[77] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. NeurIPS, 2015.

[78] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[79] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In CVPR, 2015.

[80] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. Advances in neural information processing systems, 2016.

[81] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In CVPR, 2023.

[82] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In USENIX Security Symposium, 2023.

[83] Kai Wang, Yizhou Peng, Hao Huang, Ying Hu, and Sheng Li. Mining hard samples locally and globally for improved speech separation. In ICASSP, 2022.

[84] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In ICCV, 2023.

[85] Xin Jin, Bohan Li, BAAO Xie, Wenyao Zhang, Jinming Liu, Ziqiang Li, Tao Yang, and Wenjun Zeng. Closed-loop unsupervised representation disentanglement with β-vae distillation and diffusion probabilistic feedback. arXiv preprint arXiv:2402.02346, 2024.

[86] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv preprint arXiv:2110.01963, 2021.

[87] Noa Garcia, Yusuke Hirota, Yankun Wu, and Yuta Nakashima. Uncurated image-text datasets: Shedding light on demographic bias. In CVPR, 2023.

[88] Abeba Birhane, Sanghyun Han, Vishnu Boddeti, Sasha Luccioni, et al. Into the LAION's Den: Investigating hate in multimodal datasets. In NeurIPS Datasets and Benchmarks Track, 2023.

[89] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. In NeurIPS, 2023.

[90] Eddie Ungless, Björn Ross, and Anne Lauscher. Stereotypes and smut: The (mis) representation of non-cisgender identities by text-to-image models. In ACL, 2023.

[91] Sheng-Yu Wang, Alexei A. Efros, Jun-Yan Zhu, and Richard Zhang. Evaluating data attribution for text-to-image models. In ICCV, 2023.

[92] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-Eval: Probing the reasoning skills and social biases of text-to-image generation models. In ICCV, 2023.

[93] Jialu Wang, Xinyue Gabby Liu, Zonglin Di, Yang Liu, and Xin Eric Wang. T2IAT: Measuring valence and stereotypical biases in text-to-image generation. In ACL, 2023.

[94] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. arXiv preprint arXiv:2308.00755, 2023.

[95] Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In AIES, 2023.

[96] Alexander Lin, Lucas Monteiro Paes, Sree Harsha Tanneru, Suraj Srinivas, and Himabindu Lakkaraju. Word-level explanations for analyzing bias in text-to-image models. arXiv preprint arXiv:2306.05500, 2023.

[97] Yankun Wu, Yuta Nakashima, and Noa Garcia. Stable diffusion exposed: Gender bias from prompt to image. In AIES, 2024.

[98] Yankun Wu, Yuta Nakashima, and Noa Garcia. Revealing gender bias from prompt to image in stable diffusion. Journal of Imaging, 2025.

[99] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. Communications of the ACM, 2020.

[100] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In ICML, 2016.

[101] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In ICML, 2021.

[102] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. CogView2: Faster and better text-to-image generation via hierarchical transformers. In NeurIPS, 2022.

[103] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. TMLR, 2022.

[104] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In NeurIPS, 2022.

[105] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-Prompt image editing with cross-attention control. In ICLR, 2023.

[106] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. TF-ICON: Diffusion-based training-free cross-domain image composition. In ICCV, 2023.

[107] Dave Epstein, Allan Jabri, Ben Poole, Alexei A Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. In NeurIPS, 2023.

[108] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In ICCV, 2023.

[109] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. arXiv preprint arXiv:2308.14761, 2023.

[110] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In ACL, 2023.

[111] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In ICCV, 2023.

[112] Koutilya Pnvr, Bharat Singh, Pallabi Ghosh, Behjat Siddiquie, and David Jacobs. LD-ZNet: A latent diffusion approach for text-based image segmentation. In ICCV, 2023.

[113] Abhishek Mandal, Susan Leavy, and Suzanne Little. Multimodal composite association score: Measuring gender bias in generative multimodal models. arXiv preprint arXiv:2304.13855, 2023.

[114] Hugo Berg, Siobhan Hall, Yash Bhalgat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In AACL-IJNCLP, 2022.

[115] Harvey Mannering. Analysing gender bias in text-to-image models using object detection. arXiv preprint arXiv:2307.08025, 2023.

[116] Yanzhe Zhang, Lu Jiang, Greg Turk, and Diyi Yang. Auditing gender presentation differences in text-to-image models. arXiv preprint arXiv:2302.03675, 2023.

[117] Robert Wolfe and Aylin Caliskan. American== white in multimodal language-and-image ai. In AIES, 2022.

[118] Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. ITI-GEN: Inclusive text-to-image generation. In CVPR, 2023.

[119] Melissa Hall, Candace Ross, Adina Williams, Nicolas Carion, Michal Drozdzal, and Adriana Romero Soriano. DIG In: Evaluating disparities in image generations with indicators for geographic diversity. arXiv preprint arXiv:2308.06198, 2023.

[120] Abhipsa Basu, R Venkatesh Babu, and Danish Pruthi. Inspecting the geographical representativeness of images from text-to-image models. In ICCV, 2023.

[121] Zhixuan Liu, Peter Schaldenbrand, Beverley-Claire Okogwu, Wenxuan Peng, Youngsik Yun, Andrew Hundt, Jihie Kim, and Jean Oh. SCoFT: Self-contrastive fine-tuning for equitable image generation. In CVPR, 2024.

[122] Christopher Teo, Milad Abdollahzadeh, and Ngai-Man Man Cheung. On measuring fairness in generative models. NeurIPS, 2023.

[123] Aditya Chinchure, Pushkar Shukla, Gaurav Bhatt, Kiri Salij, Kartik Hosanagar, Leonid Sigal, and Matthew Turk. TIBET: Identifying and evaluating biases in text-to-image generative models. arXiv preprint arXiv:2312.01261, 2023.

[124] Felix Friedrich, Katharina Hämmerl, Patrick Schramowski, Jindrich Libovicky, Kristian Kersting, and Alexander Fraser. Multilingual text-to-image generation magnifies gender stereotypes and prompt engineering may not help you. arXiv preprint arXiv:2401.16092, 2024.

[125] Ashutosh Sathe, Prachi Jain, and Sunayana Sitaram. A unified framework and dataset for assessing gender bias in vision-language models. arXiv preprint arXiv:2402.13636, 2024.

[126] Hanjun Luo, Haoyu Huang, Ziye Deng, Xuecheng Liu, Ruizhe Chen, and Zuozhu Liu. Bigbench: A unified benchmark for social bias in text-to-image generative models based on multi-modal llm. arXiv preprint arXiv:2407.15240, 2024.

[127] Muxi Chen, Yi Liu, Jian Yi, Changran Xu, Qiuxia Lai, Hongliang Wang, Tsung-Yi Ho, and Qiang Xu. Evaluating text-to-image generative models: An empirical study on human image synthesis. arXiv preprint arXiv:2403.05125, 2024.

[128] Yixin Wan and Kai-Wei Chang. The male ceo and the female assistant: Probing gender biases in text-to-image models through paired stereotype test. arXiv preprint arXiv:2402.11089, 2024.

[129] Wenxuan Wang, Haonan Bai, Jen-tse Huang, Yuxuan Wan, Youliang Yuan, Haoyi Qiu, Nanyun Peng, and Michael R Lyu. New job, new gender? measuring the social bias in image generation models. arXiv preprint arXiv:2401.00763, 2024.

[130] Moreno D'Incà, Elia Peruzzo, Massimiliano Mancini, Dejia Xu, Vidit Goel, Xingqian Xu, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. Openbias: Open-set bias detection in text-to-image generative models. In CVPR, 2024.

[131] Abhishek Mandal, Susan Leavy, and Suzanne Little. Generated bias: Auditing internal bias dynamics of text-to-image generative models. In ECCV workshop, 2024.

[132] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In NeurIPS, 2016.

[133] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In NeurIPS, 2017.

[134] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In CVPR, 2015.

[135] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In ACL, 2002.

[136] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, Shin ' ichiSatoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In CVPR, 2023.

[137] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV, 2014.

[138] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. In ACL, 2023.

[139] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In ACL, 2014.

[140] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In ACL, 2018.

[141] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. TextCaps: a dataset for image captioning with reading comprehension. In ECCV, 2020.

[142] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In MICCAI, 2015.

[143] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In CVPR, 2023.

[144] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159, 2024.

[145] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. arXiv preprint arXiv:2306.03514, 2023.

[146] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. In arXiv preprint arXiv:2303.05499, 2023.

[147] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In ICCV, 2023.

[148] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In EMNLP, 2017.

[149] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Gender and racial bias in visual question answering datasets. In FAccT, 2022.

[150] Steven Bird, Ewan Klein, and Edward Loper. Natural language processing with Python: analyzing text with the natural language toolkit. 2009.

[151] Robert Wolfe and Aylin Caliskan. Markedness in visual semantic ai. In FAccT, 2022.

[152] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating CLIP: towards characterization of broader capabilities and downstream implications. arXiv preprint arXiv:2108.02818, 2021.

[153] Robert Wolfe, Mahzarin R Banaji, and Aylin Caliskan. Evidence for hypodescent in visual semantic ai. In FAccT, 2022.

[154] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. In NeurIPS, 2023.

[155] Pushkar Shukla, Aditya Chinchure, Emily Diana, Alexander Tolbert, Kartik Hosanagar, Vineeth N Balasubramanian, Leonid Sigal, and Matthew Turk. Mitigate one, skew another? tackling intersectional biases in text-to-image models. arXiv preprint arXiv:2505.17280, 2025.

[156] Qiyuan He, Jinghao Wang, Ziwei Liu, and Angela Yao. Aid: Attention interpolation of text-to-image diffusion. In NeurIPS, 2024.

[157] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. In NeurIPS, 2023.

[158] Aishwarya Agarwal, Srikrishna Karanam, KJ Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. In ICCV, 2023.

[159] Jungwon Park, Jungmin Ko, Dongnam Byun, Jangwon Suh, and Wonjong Rhee. Cross-attention head position patterns can align with human visual concepts in text-to-image generative models. In ICLR, 2025.

[160] Jeeyung Kim, Erfan Esmaeili, and Qiang Qiu. Text embedding is not all you need: Attention control for text-to-image semantic alignment with text self-attention maps. arXiv preprint arXiv:2411.15236, 2024.

[161] Yue Jiang, Yueming Lyu, Ziwen He, Bo Peng, and Jing Dong. Mitigating social biases in text-to-image diffusion models via linguistic-aligned attention guidance. In ACM MM, 2024.

[162] Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. In ICCV, 2023.

[163] Dvir Samuel, Rami Ben-Ari, Nir Darshan, Haggai Maron, and Gal Chechik. Norm-guided latent space exploration for text-to-image generation. In NeurIPS, 2023.

[164] Hu Yu, Hao Luo, Fan Wang, and Feng Zhao. Uncovering the text embedding in text-to-image diffusion models. arXiv preprint arXiv:2404.01154, 2024.

[165] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. TOG, 2022.

[166] Xudong Shen, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, and Mohan Kankanhalli. Finetuning text-to-image diffusion models for fairness. In ICLR, 2024.

[167] Jia Li, Lijie Hu, Jingfeng Zhang, Tianhang Zheng, Hua Zhang, and Di Wang. Fair text-to-image diffusion via fair mapping. In AAAI, 2025.

[168] Piero Esposito, Parmida Atighehchian, Anastasis Germanidis, and Deepti Ghadiyaram. Mitigating stereotypical biases in text to image generative systems. arXiv preprint arXiv:2310.06904, 2023.

[169] Hidir Yesiltepe, Kiymet Akdemir, and Pinar Yanardag. Mist: Mitigating intersectional bias with disentangled cross-attention editing in text-to-image diffusion models. arXiv preprint arXiv:2403.19738, 2024.

[170] Junlei Zhou, Jiashi Gao, Xiangyu Zhao, Xin Yao, and Xuetao Wei. Association of objects may engender stereotypes: Mitigating association-engendered stereotypes in text-to-image generation. In NeurIPS, 2024.

[171] Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In CVPR, 2024.

[172] Yilei Jiang, Weihong Li, Yiyuan Zhang, Minghong Cai, and Xiangyu Yue. Debiasdiff: Debiasing text-to-image diffusion models with self-discovering latent attribute directions. arXiv preprint arXiv:2412.18810, 2024.

[173] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. arXiv preprint arXiv:2302.00070, 2023.

[174] Colton Clemmer, Junhua Ding, and Yunhe Feng. Precisedebias: An automatic prompt engineering approach for generative ai to mitigate image demographic biases. In WACV, 2024.

[175] Min Hou, Yueying Wu, Chang Xu, Yu-Hao Huang, Chenxi Bai, Le Wu, and Jiang Bian. Invdiff: Invariant guidance for bias mitigation in diffusion models. In KDD, 2025.

[176] Rishubh Parihar, Abhijnya Bhat, Abhipsa Basu, Saswat Mallick, Jogendra Nath Kundu, and R Venkatesh Babu. Balancing act: distribution-guided debiasing in diffusion models. In CVPR, 2024.

[177] Mintong Kang, Vinayshekhar Bannihatti Kumar, Shamik Roy, Abhishek Kumar, Sopan Khosla, Balakrishnan Murali Narayanaswamy, and Rashmi Gangadharaiah. Fairgen: Controlling sensitive attributes for fair generations in diffusion models via adaptive latent guidance. arXiv preprint arXiv:2503.01872, 2025.

[178] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In CVPR, 2023.

[179] Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. arXiv preprint arXiv:2302.10893, 2023.

[180] Jinya Sakurai and Issei Sato. Fairt2i: Mitigating social bias in text-to-image generation via large language model-assisted detection and attribute rebalancing. arXiv preprint arXiv:2502.03826, 2025.

[181] Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, and Vasudev Lal. Socialcounterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In CVPR, 2024.

[182] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, 2018.

List of Publications

Journal Publications (related to this thesis)

1. **Yankun Wu**, Yuta Nakashima, and Noa Garcia. Revealing Gender Bias from Prompt to Image in Stable Diffusion. Journal of Imaging. Vol.11, Issue 2, No.35. 2025.

2. **Yankun Wu**, Yuta Nakashima, and Noa Garcia. GOYA: Leveraging Generative Art for Content-Style Disentanglement. Journal of Imaging. Vol.10, Issue 7, No.156. 2024.

International Conference (related to this thesis)

1. **Yankun Wu**, Yuta Nakashima, and Noa Garcia. Stable diffusion exposed: Gender bias from prompt to image. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. Vol. 7. pp. 1648-1659. 2024.

2. **Yankun Wu**. Generative Models for Art and Society. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Vol. 7, No. 2, pp. 58-60. 2024.

3. **Yankun Wu**, Yuta Nakashima, Noa Garcia, Sheng Li, and Zhaoyang Zeng. Reproducibility Companion Paper: Stable Diffusion for Content-Style Disentanglement in Art Analysis. Proceedings of the 2024 International Conference on Multimedia Retrieval, pp. 1228-1231. 2024.

4. **Yankun Wu**, Yuta Nakashima, and Noa Garcia. Not only generative art: Stable dif-

fusion for content-style disentanglement in art analysis. Proceedings of the 2023 ACM International conference on multimedia retrieval. pp. 199-208. 2023.

Journal Publications (not related to this thesis)

1. Yuta Nakashima, Yusuke Hirota, **Yankun Wu**, and Noa Garcia. Societal Bias in Vision-and-Language Datasets and Models. NIHON GAZO GAKKAISHI (Journal of the Imaging Society of Japan). Vol.62, No. 6, pp. 599-609, 2023.

International Conference (not related to this thesis)

1. **Yankun Wu**, Zakaria Laskar, Giorgos Kordopatis-Zilos, Noa Garcia, Giorgos Tolias. Instance-Level Generation for Representation Learning. International Conference on Computer Vision. 2025. (submitted)

2. Rawisara Lohanimit, **Yankun Wu**, Amelia Katirai, Yuta Nakashima, Noa Garcia. Privacy in Image Datasets: A Case Study on Pregnancy Ultrasounds. AAAI/ACM Conference on AI, Ethics, and Society. 2025. (submitted)

3. **Yankun Wu**, Yuta Nakashima, and Noa Garcia. Gender Bias Evaluation in Text-to-image Generation: A Survey. Critical Evaluation of Generative Models and Their Impact on Society Workshop (CEGIS Workshop) at European Conference on Computer Vision. 2024.

4. Noa Garcia, Yusuke Hirota, **Yankun Wu**, and Yuta Nakashima. Uncurated image-text datasets: Shedding light on demographic bias. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6957-6966. 2023.

Fellowship

1. Yankun Wu, 次世代挑戦的研究者育成プロジェクト. (2024 - 2025)