Self-Discovering Interpretable Diffusion Latent Directions for Responsible Text-to-Image Generation

Hang Li[1,4,5] Chengzhi Shen[3] Philip Torr[2] Volker Tresp[1,4] Jindong Gu[2*]
[1]LMU Munich, Germany [2]University of Oxford, UK [3]Technical University of Munich, Germany
[4]Munich Center for Machine Learning, Germany [5]Siemens AG, Germany

Abstract

Diffusion-based models have gained significant popularity for text-to-image generation due to their exceptional image-generation capabilities. A risk with these models is the potential generation of inappropriate content, such as biased or harmful images. However, the underlying reasons for generating such undesired content from the perspective of the diffusion model's internal representation remain unclear. Previous work interprets vectors in an interpretable latent space of diffusion models as semantic concepts. However, existing approaches cannot discover directions for arbitrary concepts, such as those related to inappropriate concepts. In this work, we propose a novel self-supervised approach to find interpretable latent directions for a given concept. With the discovered vectors, we further propose a simple approach to mitigate inappropriate generation. Extensive experiments have been conducted to verify the effectiveness of our mitigation approach, namely, for fair generation, safe generation, and responsible text-enhancing generation. Project page: https://interpretdiffusion.github.io.

1. Introduction

The rapid advances in vision language models have sparked increasing interest in ensuring their safety and responsible use [7, 23, 24]. In particular, text-to-image diffusion models, which have exhibited remarkable performance in creating images from text prompts [13, 15, 20, 31, 35, 37, 49], raise concerns about the risks of generating inappropriate content. The generated images may exhibit biases and unsafe elements, including instances of gender discrimination or the depiction of violent scenes that could be harmful to children. Recent research efforts have focused on introducing safety mechanisms to mitigate these issues, such as filtering out inappropriate text input, detecting inappropriate images with a safety guard classifier [4, 32, 36, 38] and building safe diffusion models [5, 6, 17]. However, the underlying mechanism of how diffusion models generate inappropriate content remains poorly understood. In this work, we aim to explore the following questions. 1) Are there any internal representations associated with these inappropriate concepts in the diffusion model-based generation process? 2) Can we manipulate representations to avoid inappropriate content corresponding to a given concept, i.e., to achieve responsible image generation?

To understand the image generation process of diffusion models, previous work has identified the bottleneck layer of the U-Net as a semantic representation space, dubbed h-space [18]. They demonstrated that a vector in the h-space can be associated with a specific semantic concept in the generated image. Manipulating the vector in the space can alter the generated image in a semantically meaningful way, such as adding a smile to a face. Several approaches [9, 18, 30] have been proposed to discover meaningful directions in this h-space. For instance, an approach in [9] uses PCA to identify a set of latent directions that may represent semantic concepts.

However, existing approaches to identifying interpretable latent vectors are limited. In unsupervised approaches [9, 30], it is not clear to which semantic concepts those identified vectors correspond. The found vectors must be interpreted with humans in a loop. Furthermore, the number of interpretable directions depends on the training data [9, 30]. It is highly likely that some target concepts may not be found in the discovered directions, especially those related to fairness and safety. Supervised approaches [9, 18] have also been explored to identify target concepts. These methods require training external attribute classifiers supervised by human annotations. Additionally, the quality of the identified vectors is sensitive to the classifier's performance. Furthermore, new concepts require the training of new classifiers. Overall, existing interpretation methods cannot be easily applied to identify the corresponding semantic vector for a given inappropriate concept.

In this work, we propose a self-discovery approach to find interpretable latent directions in the h-space for user-

*Corresponding author

defined concepts. We learn a latent vector that effectively represents the concept by leveraging the model's acquired semantic knowledge in its internal representations. Initially, images are generated using specific text prompts related to the concept. The images are then used in a denoising process where the frozen pretrained diffusion model reconstructs these images from noise, guided by a modified text prompt that omits the desired concept, and our introduced latent vector. By minimizing the reconstruction loss, the vector learns to represent the given concept. Our self-discovery approach eliminates the need for external models like CLIP text encoder [34] or dedicated attribute classifiers trained on human-labeled datasets. We identify ethical-related latent vectors and demonstrate their applications in responsible text-to-image generation: 1) fairness by sampling an ethical concept, e.g., gender, in the latent space, which generates images with unbiased attributes and aligned with the prompt. 2) safety generation by incorporating safety-related concepts, e.g., one that eliminates the nudity content, into h-space to prevent the model from generating such harmful content. 3) responsible guidance, where we first discover responsible concepts in the text prompt and enhance the expression of those ethical concepts.

Previous approaches enhance responsible image generation from different perspectives. Concretely, [2, 5, 6, 8, 17] fine-tune the diffusion models or text embeddings to unlearn harmful concepts, and [39] applies classifier-free guidance to steer the generation away from unsafe concepts. Despite the mitigation mechanism of previous approaches, diffusion models still suffer from inappropriate content generation [39, 50]. Unlike previous work, in this work, we provide a new perspective to mitigate the inappropriate generation, namely, finding and manipulating concepts in an interpretable latent space. Our work can be easily combined with previous mitigation approaches to further enhance responsible text-to-image generation.

We conducted extensive experiments on fairness, safety, and responsible guidance-enhancing generation. Our model consistently produces images with a balanced representation across societal groups. Further, we successfully mitigate harmful content for inappropriate prompts. In addition, our approach synergistically improves the performance of responsible image generation when combined with existing methods. Furthermore, we enhance text guidance to generate fair and safe content for responsible prompts.

Our contributions can be summarized as follows:

- We propose a self-discovery method for identifying interpretable directions in the diffusion latent space. Our approach can find a vector that represents any desired concept, without the need for labeled data or external models.
- With the discovered vectors, we propose a straightforward yet effective approach to enhance responsible generation, including fair generation, safe generation, and responsible text-enhancing generation.
- Extensive experiments are conducted to validate the effectiveness of our approach.

2. Related Work

Responsible Alignment of Diffusion Models Various approaches have been proposed to mitigate the generation of biased and unsafe content in diffusion models. A straightforward method involves refining the training dataset to remove biased and inappropriate content, exemplified by Stable Diffusion (SD) v2 [37]. Such approaches can be computationally intensive, may not fully eliminate harmful content [5], and could degrade the model's performance [39]. An alternative is to detect and filter out inappropriate words from the input prompts [1, 2, 27]. However, this fails to address non-explicit phrases that can still yield inappropriate outputs. Another line of approaches involves finetuning the parameters of pretrained models, aiming to remove the model's representation capability of generating such inappropriate concepts [5, 17]. However, they are sensitive to the adaptation process and may result in the degradation of the original models [6, 10, 28, 29, 48]. Moreover, such approaches require a potentially exhaustive list of words that introduce biases and harmful concepts [5, 6, 29]. Training-free approaches utilize classifier-free guidance to direct the generated images away from undesirable content during inference [1, 3, 39, 39, 47]. While they modify the noise space using text-based guidance through cross-attention mechanisms, we adopt a similar conditioning strategy to manipulate the generation for frozen pretrained models in the semantic latent space. As an orthogonal approach to the existing literature, we mitigate the inappropriate content by finding the corresponding latent directions in the U-Net bottleneck layer and suppressing their activations.

Interpreting Diffusion Models To understand the working mechanisms of diffusion models, recent works mainly focus on investigating text guidance for conditional diffusion models [11, 16, 22, 26, 29, 43, 46], or analyzing the internal representations in diffusion models' intermediate layer activations [9, 18, 30, 40, 44]. We focus on elucidating the internal representations learned within the diffusion model, in line with prior works [18]. Some work [19, 33, 45, 51] proposes to create a semantic space in diffusion models by employing an autoencoder to encode the image into a semantic vector that guides the decoding process. However, their approaches require adapting the parameters of the autoencoder or even the entire framework.

The seminal work [18] reveals that the bottleneck layer of U-Net architecture already exhibits properties suitable for a semantic representation space. They identified disentangled representations associated with the semantics of the generated image and demonstrated that those latent di-

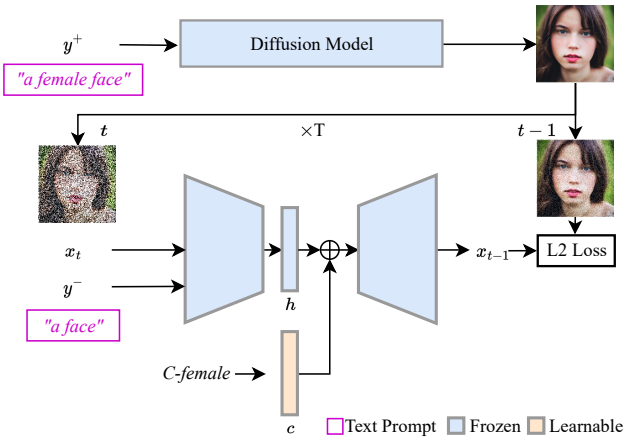

Figure 1. Optimization framework to discover a semantic vector for a given concept. The top line shows that an image is firstly generated by the pretrained Stable Diffusion model for the prompt "a female face". The bottom part shows the optimization process for finding the concept for "female" in the semantic h-space. The concept vector is used to reconstruct the image along with a modified prompt "a face", under an iterative denoising process. With the pretrained diffusion model frozen, the gradients of the reconstruction loss can solely update the latent vector to represent the missing gender information. After convergence, the latent vector is aligned with the U-Net's internal representation of the "female" concept, which can be used to guide new image generation.

rections are identical to different images. However, their approach relies on the CLIP classifier and paired source-target images and edits, making it inefficient. Another work proposes a PCA-based decomposition method on the latent space and finds interpretable attribute directions using the top right-hand singular vectors of the Jacobian. Additionally, [30] uses Riemannian metrics to define more accurate and meaningful directions. However, these approaches require manual interpretation to identify the editing effect of each component. Our approach differs from the supervised approach in [9] by enabling the efficient discovery of latent directions for any given target concept without requiring a data collection process or training external classifiers.

3. Approach

This section first introduces our optimization method to find interpretable directions in diffusion models' h-space. In the second part, we show how to utilize discovered concepts in the inference process for responsible generation, including fairness, safety, and text-enhancing generation.

3.1. Finding a Semantic Concept

Diffusion models are generative models that generate samples from Gaussian noise through a denoising process [13, 41, 42]. Starting from a random vector $x_T \sim \mathcal{N}(0,1)$ of the same dimension as the image, the model estimates a noise value at each time step to subtract from the current vector to obtain a denoised image, denoted as $x_{t-1} = x_t - \epsilon_\theta(x_t, t)$, where ϵ_θ represents the U-Net of the diffusion model. A clean image x_0 is obtained at the end of this denoising process. The training of diffusion models involves a forward process that iteratively adds noise to images from the data, denoted as $x_t = x_{t-1} + \epsilon_t$, with $\epsilon_t \sim \mathcal{N}(0,1)$. The training loss includes predicting noises for different steps,

$$L = \sum_{x\sim\mathcal{D}} \sum_{t\sim[0,T]} \|\epsilon - \epsilon_\theta(x_t, t)\|^2. \quad (1)$$

Recent work identified a semantic space in the diffusion model, the activations of U-Net's bottleneck layer h, as shown in Figure 1. The activations in h-space leads to the generation of a less noised image for the next timestep x_{t-1}[1]. This space exhibits semantic structures and is easy to interpret. Activating a specific vector in the U-Net bottleneck layer leads to the image having a certain attribute. However, existing approaches cannot find the vector for an arbitrarily given concept. Our goal is to find such vectors.

To this end, we utilize the text-to-image conditional diffusion model which can generate images from a given text input. The prediction function in Eq. 1 becomes $\epsilon_\theta(x_t, \pi(y), t)$ where $\pi(y)$ is the encodings of the input text y. The equation specifies a conditional distribution that drives the generation of the image towards data regions that are highly likely given the input text [12]. To discover an interpretable direction, we leverage the pre-trained model to generate a set of images using dedicated prompts related to that concept. For example, to find the latent direction of the concept "female", we first generate a set of images x^+ with a descriptive prompt y^+ "a photo of a female face". Then, a concept vector is optimized for the conditional generation where the original prompt has been modified into y^- "a photo of a face", eliminating gender information. The concept vector $c \in \mathbb{R}^D$ is randomly initialized in the latent space, where D is the dimension of h-space, and is optimized to minimize the reconstruction error. Since the pre-trained diffusion model is frozen, the model has to utilize the extra condition c to compensate for the missing information not in the text condition but in the image. The concept vector c will be forced to represent the missing information from the input text to produce an image with the lowest reconstruction error. After convergence, that vector c is expected to represent the gender information "female". In this way, we discover a set of vectors that represent target concepts, such as gender, safety, and facial expressions.

[1]The decoding of x_{t-1} depends on other variables due to the presence of skip connections. For simplicity, we omit this consideration as the skip connection seems less significant in encoding compact semantic information, as supported in previous findings [9, 14].

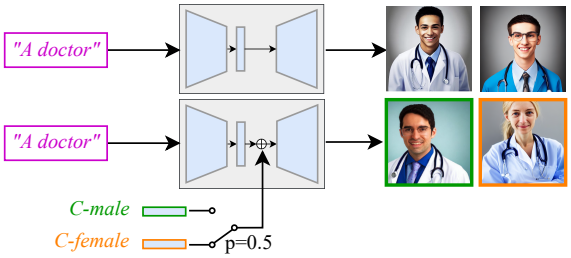

Figure 2. Fair Generation. Top: images generated from the prompt "doctor" are biased toward males. Bottom: we sample a learned male or female concept with equal probability for generating the doctors. The doctors now have fair gender. Images are generated from different random seeds.

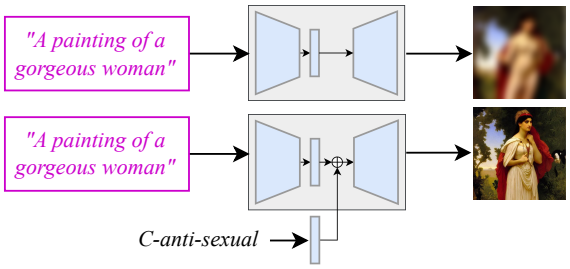

Figure 3. Safe Generation. When the user's prompt contains implicit references to nudity, the original model (shown in the top row) generates an inappropriate image, as the added blurriness indicates. In contrast, our approach generates an image for the same prompt by setting a safety-related concept in h-space, identified in the previous section. The vector anti-sexual concept represents the direction to suppress nudity content, effectively eliminating inappropriate content while maintaining fidelity to the prompt.

Formally, the optimal c^* for a given concept is found by

$$c^* = \arg\min_{c} \sum_{x,y\sim\mathcal{D}} \sum_{t\sim[0,T]} \|\epsilon - \epsilon_\theta(x_t^+, t, \pi(y^-), c)\|^2, \quad (2)$$

x_t^+ denotes the noised version of the original image generated with y^+, c represents the target concept. ϵ_θ denotes the U-Net that linearly adds an additional concept vector c to its h-space, at each decoding timestep. Regarding implementation, the h-space is the flattened activations after the middle bottleneck layer of the U-Net. The pseudo-code for this training pipeline is in Appendix A.1.

We learn a single vector for each concept for all timesteps, as the latent direction remains approximately consistent across different timesteps [18]. Moreover, we restrict the operation to linearity to demonstrate the power of this latent space. Notably, the learned vector generalizes effectively to new images [9] and diverse prompts [30]. For instance, a "male" concept learned with the base prompt "person" can be used in different contexts, such as "doctor" or "manager", as shown in the next section. Additionally, the concepts can be optimized jointly or independently, with the experimental section demonstrating the impact of concept composition. A key strength of our approach is utilizing synthesis by diffusion models to collect data, eliminating the need for human labeling and training of guiding classifiers. Nevertheless, our approach can be applied to realistic datasets with annotated attributes.

3.2. Responsible Generation with Self-discovered Interpretable Latent Direction

In this subsection, we utilize the identified directions to manipulate the latent activation in the latent space for fair, safe, and enhanced responsible generation.

Fair Generation Method A text prompt contains words that lead to the generation of biased societal groups. We aim to generate images with evenly distributed attributes for a given text prompt. For example, for the prompt "doctor", we aim to generate an image of a male doctor with a 50% probability and a female doctor with a 50% probability. For that, we learn a set of semantic concepts representing different societal groups using the approach in the previous section. For inference, a concept vector is sampled from the learned concepts in the societal group with equal probability, e.g., the C-male and C-female concept vectors for gender are chosen with fair chance. The inference process is fixed as before, except that the sampled vector is added to the original activations in h-space at each decoding step, denoted by

$$h \leftarrow h + c \sim \text{Categorical}(p_k), \quad (3)$$

where p_k represents the probability of sampling a particular attribute c_k from the societal group with C distinct attributes. For the fair generation, $p_k = 1/C$. Guided by this sampled concept vector, the generated image is expected to be a male doctor if the C-male concept is sampled or a female doctor otherwise. This allows the generated images to have an equal number of attributes, e.g., an equal number of male and female doctors, shown in Figure 2.

Safe Generation Method For safety generation, we consider text prompts that contain explicit or implicit references to inappropriate content, which we aim to eliminate. An example of such a prompt is illustrated in Figure 3, where the phrase "a gorgeous woman" may indirectly lead to the generation of nudity. We identify a collection of safety-related concepts, such as anti-sexual, to achieve safe generation.

Specifically, we learn the opposite latent direction of an inappropriate concept, leveraging the negative prompt technique. For instance, the training images are generated by the prompt y^+ "a gorgeous person" with a negative prompt "sexual", which effectively instructs the Stable Diffusion to generate safe images without sexual content. The concept vector is then optimized on those training images that depict

safe content. For that, the input prompt y^- is set to "a gorgeous person" but without the negative prompt "sexual". In this way, the concept vector directly learns the concept of "anti-sexual". The reason for adopting this strategy is the difficulty of listing all the opposite concepts of sexuality, e.g., "dressed", "clothes", or more. An alternative approach is to learn the concept of sexuality directly and apply negation during generation, which we found less effective. More details regarding the negative prompt are in Appendix A.2.

After the learning process, we maintain all aspects of inference unchanged, except for adding the learned vector to the original activations at the bottleneck layer, formally as

$$h \leftarrow h + c_s. \tag{4}$$

Here, c_s refers to a safety concept, such as "anti-sexual", which represents the opposite of sexual content. This strengthens the expression of safe concepts in the generated images so they are devoid of harmful content. Figure 3 illustrates the impact of including the anti-sexual vector, resulting in a visually appealing person with appropriate clothes.

Responsibile Text-enhancing Generation Method Even when a prompt is intentionally designed to promote safety, the generative models may struggle to accurately incorporate all the concepts defined in the prompt. For instance, consider a text prompt like "an exciting Halloween party, no violence". The generative model may encounter difficulties in faithfully representing each responsible concept from the prompt, e.g., its poor understanding of negation on "violence" may result in the generation of inappropriate content.

To address this issue, we utilize our self-discovery approach to learn concepts such as gender, race, and safety. To enhance the generation of responsible prompts, we extract safety-related content from the text and leverage our learned ethical-related concepts to reinforce the expression of desired visual features. During inference, we apply the extracted concepts $c(y)$ from the prompt to the original activations, denoted as

$$h \leftarrow h + c(y) \tag{5}$$

For example, as shown in Figure 4, the concept of "no violence" from the text prompt activates our learned "anti-violence" concept during inference. By directly manipulating the semantic space, our approach introduces the desired attributes to the generated image. Compared to the original generated image, the anti-violence concept effectively mitigates the presence of violent content and makes the generated images more appropriate.

4. Experiments

In this section, we present the performance of our methods on fair generation, safe generation, and enhanced responsible generation. Additionally, we demonstrate the general properties of our discovered concept vectors, highlighting their generalization capability and compositionality.

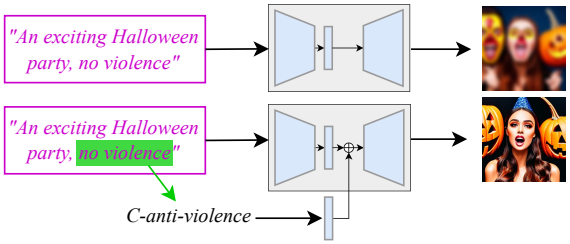

Figure 4. Responsible text-enhancing generation. The original model may fail to capture the safety concepts specified in the text, such as "no violence". We propose extracting those safety concepts from the given prompt and activating the safety directions during generation. The bottom image demonstrates that incorporating our safety concepts can enhance the text guidance of the original prompt.

4.1. Fair Generation

Task The goal is to increase the diversity of societal groups in the generated images, particularly in professions where existing models exhibit gender and racial bias.

Dataset Following [6, 29], our approach is evaluated on the Winobias [52] benchmark that comprises 36 professions known to exhibit gender biases. We generate five different prompts for each profession using various templates, such as "a portrait of a doctor". Additionally, we create a hard dataset by augmenting the existing dataset with prompts that are more likely to induce stereotypes. This extended dataset addes the term "successful" to each original prompt, such as "a portrait of a successful doctor". The word "successful" often leads to the generation of male images and thus constitutes challenges for a fair generation [6]. Appendix B.1 contains the complete list of prompts.

Evaluation Metric The CLIP classifier is employed to predict attributes by measuring the similarity between the text embedding of a concept (e.g., female, male) and the embedding of the generated image. We utilize the deviation ratio [6, 29] to quantify the imbalance of different attributes. To accommodate an arbitrary number of attributes, the metric is modified as $\Delta = \max_{c \in C} \frac{|N_c/N - 1/C|}{1 - 1/C}$, where C is the total number of attributes within a societal group, N is the total number of generated images, and N_c denotes the number of images whose maximum predicted attribute equals c. In particular, we test the gender, *male*, *female*, and racial, *black*, *white*, *Asian*, biases associated with the professions. These races are selected as the CLIP classifier has relatively reliable predictions on these attributes. During the evaluation, 150 images were generated for each profession.

Approach Setting In all experiments, we use the Stable Diffusion v1.4 checkpoint and set the guidance scale to 7.5

(a) Images generated by **SD**

(b) Images generated by **Ours**

Figure 5. Gender fairness generation. From the prompt "a photo of a doctor", the original SD exhibits significant gender bias, as shown on the left side. Our approach with uniformly sampled gender vectors represents genders equally in the generated images.

Dataset	Gender			Gender+			Race			Race+		
Method	SD	UCE	Ours	SD	UCE	Ours	SD	UCE	Ours	SD	UCE	Ours
Analyst	0.70	0.20	**0.02**	0.54	0.04	**0.02**	0.82	0.29	**0.24**	0.77	**0.20**	0.41
CEO	0.92	0.28	**0.06**	0.90	0.58	**0.06**	0.38	**0.13**	0.22	0.31	**0.08**	0.22
Laborer	1.00	**0.09**	0.12	0.98	**0.08**	0.14	0.33	0.40	**0.24**	0.53	0.38	**0.20**
Secretary	0.64	**0.10**	0.36	0.92	0.96	**0.46**	0.37	0.35	**0.24**	0.55	0.43	**0.32**
Teacher	0.30	0.06	**0.04**	0.48	0.16	**0.10**	0.51	0.10	**0.04**	0.26	0.23	**0.21**
Winobias[52]	0.68	0.22	**0.17**	0.70	0.52	**0.23**	0.56	**0.21**	0.23	0.48	0.35	**0.20**

Table 1. Fair generation quantified by the deviation ratio ($0 \leq \Delta \leq 1$). Lower values indicate better performance. The left side of the table presents the results on gender attributes, whereas the right side quantifies the racial bias. "Gender+/Race+" refers to the extended Winobias dataset, which is more challenging, as described in Subsection 4.1. Our approach leads to unbiased generation for biased prompts and is robust to diverse sources of bias in the prompt.

for text-to-image generation. We find five concept vectors using a base prompt "person", e.g., y^+ = "a photo of a woman" and y^- = "a photo of a person" to learn the concept "female". The concept vectors are optimized for 10K steps on 1K synthesized images for each concept. During inference, we directly employ the learned vector without any scaling. Unlike the baseline approach UCE [6], which needs to debias each profession in Winobias, Our approach is trained solely on the "person" prompt to learn the male and female concept that generalizes to all different professions. For comparison, we report UCE's published scores when available and otherwise use their released code to train the model.

Results and Analysis Table 1 reveals that our approach is significantly better than the original SD and outperforms the state-of-the-art debiasing approach UCE. The professions on the table are randomly selected from the complete list of 36 professions (see Appendix B.2). Figure 5 compares images generated from our approach and those from the original SD. Further, we highlight the generalization capability of our approach to different text prompts using the extended Winobias dataset, as shown in the second and fourth column blocks of Table 1. Despite the presence of bias in the text prompts, our approach consistently performs well as it directly operates on the latent visual space. In contrast, UCE performs poorly on this challenging dataset, as it relies on debiasing each word in the prompt. The effectiveness of UCE is easily weakened by biased words that are not included in its training set. Additionally, we demonstrate that the quality of images generated by our approach remains consistent with the original SD and UCE in Appendix B.3.

4.2. Safe Generation

Task This section focuses on generating images that eliminate harmful content specified in inappropriate prompts. As an orthogonal approach to existing methods, our approach is combined with current safety methods, including SLD [39] and ESD [5], to eliminate inappropriate generation further.

Dataset and Evaluation Metric The I2P benchmark [39] is a collection of 4703 inappropriate prompts from real-world user prompts. The inappropriateness covers seven categories, including, e.g., illegal activity, sexual, and violence. For evaluation, the Nudenet[2] detector and Q16 [38] classifier are used to detect nudity or violent content in an image. An image is classified as inappropriate if any of the classifiers predicts a positive [5]. Five images are generated for each prompt for evaluation.

Approach Setting We find that optimizing a single concept vector for "safety" is challenging. Therefore, we learn the concept vector for each inappropriate concept defined in the

[2] https://github.com/notAI-tech/NudeNet

Category	Harassment	Hate	Illegal	Self-harm	Sexual	Shocking	Violence	I2P[39]
SD	0.34 ±0.019	0.41 ±0.032	0.34 ±0.018	0.44 ±0.019	0.38 ±0.016	0.51 ±0.017	0.44 ±0.018	0.41 ±0.007
Ours-SD	**0.18** ±0.015	**0.29** ±0.030	**0.23** ±0.016	**0.28** ±0.017	**0.22** ±0.014	**0.36** ±0.017	**0.30** ±0.017	**0.27** ±0.006
SLD[39]	0.15 ±0.014	**0.18** ±0.025	0.17 ±0.015	0.19 ±0.015	0.15 ±0.012	0.32 ±0.016	0.21 ±0.015	0.20 ±0.006
Ours-SLD	**0.14** ±0.014	0.20 ±0.027	**0.14** ±0.013	**0.14** ±0.013	**0.09** ±0.010	**0.25** ±0.015	**0.16** ±0.013	**0.16** ±0.005
ESD[5]	0.27 ±0.018	0.32 ±0.031	0.33 ±0.018	0.35 ±0.018	0.18 ±0.013	**0.41** ±0.017	0.41 ±0.018	0.32 ±0.007
Ours-ESD	**0.26** ±0.017	**0.29** ±0.030	**0.25** ±0.017	**0.26** ±0.017	**0.13** ±0.011	**0.41** ±0.017	**0.30** ±0.017	**0.27** ±0.006

Table 2. The proportion of images classified as inappropriate on the I2P benchmark. In each block of results, the first row shows the performance of the original method, while the second row represents adding our concept vector to the corresponding baseline model. Our identified safety-related vector can be combined with existing safety approaches to mitigate inappropriate content generation.

I2P dataset, e.g., "anti-sexual". The identified concept vectors are linearly combined as the final vector to modulate the generation process. The effect of linearly combining concept vectors is further discussed in the next section. Additionally, certain concepts are rather abstract and include diverse visual categories, e.g., "hate". Adding these concepts improves safety yet at a higher cost of image quality degradation. Therefore, we utilize two prevailing concepts, anti-sexual and anti-violence, which may underline other concepts, as our final model. For baselines, we use SLD [39] and ESD [5] with their released model weights. Detailed settings and hyperparameters are in Appendix C.

Results and Analysis Table 2 demonstrates the effectiveness of our approach in eliminating inappropriate content. Notably, our safety vector can suppress inappropriate concepts that existing approaches failed to eliminate, as evidenced by the improved performance compared to existing approaches. In particular, we observe significant improvement in eliminating nudity with the combined approach of SLD and our proposed method, with a relative improvement of 40%. Additionally, even after retraining the model weights to eliminate nudity, ESD still generates inappropriate images, which our approach can mitigate.

4.3. Enhancing Responsible Text Guidance

Task For user prompts classified as responsible text, we aim to accurately represent the responsible phrases in the prompt in the generated image.

Dataset We created a dataset of 200 prompts that explicitly include responsible concepts. In particular, the dataset focuses on gender and race fairness, as well as the removal of sexual and violent content. For instance, the prompt "a fair-gender doctor is operating a surgery" assesses the model's ability to generate images with balanced genders based on the text cue "fair-gender". Additionally, prompts such as "a picture of a loved couple, without sexual content" address safety concerns. More dataset details are in Appendix D.

Results and Analysis We utilized previously introduced metrics for evaluation, i.e., the deviation ratio for fairness, NudeNet for sexual content, and Q16 for violent content.

Model	Gender	Race	Sexual	Violence
SD	0.2199	0.1600	0.4300	0.4551
Ours-SD	**0.1433**	**0.1399**	**0.2640**	**0.3204**

Table 3. For prompts containing responsible concepts, the original SD may fail to follow the prompts faithfully. Our approach effectively enhances responsible text-guidance generation.

For evaluation, 3500 images are generated for the dataset. For our approach, we provide the model with the corresponding concept associated with the input prompt. For example, if a prompt mentions "no sexual", the anti-sexual vector is added to the generation process. Table 3 compares our approach with the original model, which does not use the safety concepts. Our approach effectively enhances the text guidance for responsible instructions.

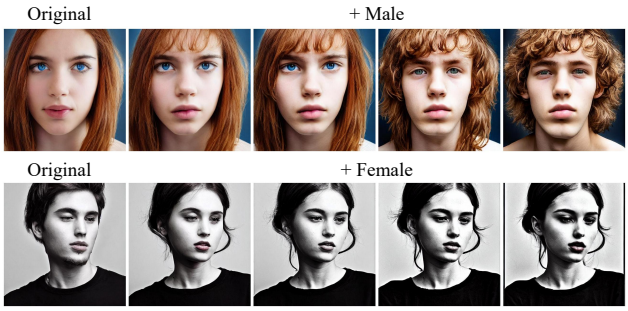

Figure 6. Concept interpolation. The first column displays the original image generated by SD. The following columns show images generated by the same random seeds, but with concept vector scales linearly increasing from 0.2 to 0.8.

4.4. Semantic Concepts

In previous experiments, we have demonstrated the specific applications of our identified concept vectors for responsible generation. This subsection introduces the general properties of discovered vectors related to the semantic space.

Interpolation Figure 6 illustrates the impact of manipulating image semantics by linearly controlling the strength of the concept vector, denoted as λ in the equation $h \leftarrow h + \lambda c$.

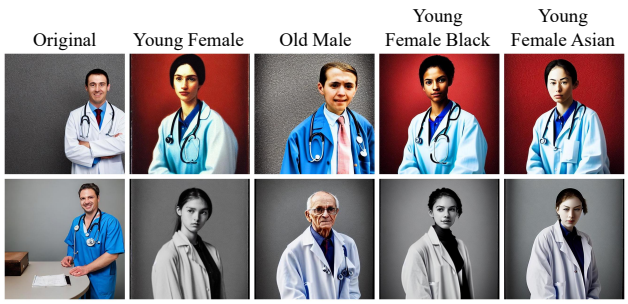

Figure 7. Multiple concepts composition. The concept vectors of gender, age, and race were learned independently. Linearly adding latent vectors can generate images with corresponding semantics.

The image is gradually modified to the introduced concept by adjusting the added vector's strength. The smooth transition indicates that the discovered vector represents the target semantic concept while remaining approximately disentangled from other semantic factors. Appendix E.1 presents more examples of concept manipulation and enhanced fidelity by post-hoc interpolation methods [25].

Composition Figure 7 showcases the composability of learned concept vectors, which were trained independently. Images are generated from the prompt "a photo of a doctor". By linearly combining these concept vectors, we can control the corresponding attributes in the generated image. Appendix E.2 provides a quantitative evaluation.

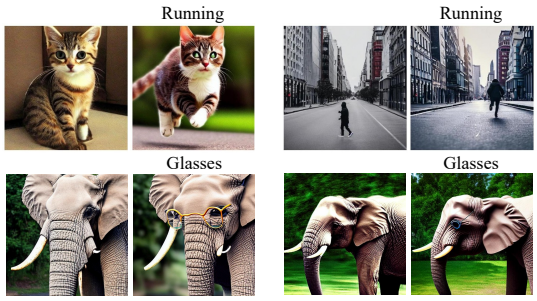

Figure 8. General semantic concepts identified by our approach. Top: The concept of "running" is learned from dog images and can be generalized to different objects. Bottom: The concept vector of "glasses" enhances the prompt "an elephant wearing glasses".

Generalization Figure 8 illustrates the generalization capability of our discovered concept vector to universal semantic concepts. We train the latent vector for the concept "running" on generated dog images and test its effect on other objects using prompts such as "a photo of a cat". Each pair of images in the figure is generated from the same random seed. Although the vector of "running" was learned from dogs, it successfully extends to different animals and even humans. Additionally, our approach enhances the original text guidance for the prompt "an elephant wearing glasses". The original SD cannot produce accurate images, as shown in the first and third images on the bottom. The correct images can be generated by adding the concept vector "glasses" in h-space. More visualizations are in Appendix E.3.

Model	SD	SLD*	SLD	ESD*	ESD	Ours-SD
FID	14.09	16.90	19.35	13.68	15.36	15.98
CLIP	31.33	–	30.41	–	30.05	31.03

Table 4. Evaluation of the quality of generated images on the COCO-30K [21] dataset using FID for image fidelity and CLIP Score for semantic alignment with input text. Various safety approaches have approximately the same level of image quality as the original SD. Numbers reported from the corresponding papers are denoted with *.

Impact on Image Quality Additionally, we find that the quality of generated images remains approximately the same level as the original SD, as shown in Table 4. The observed differences in the reported scores and these in our experiments can be attributed to the randomness during image generation and caption sampling, which aligns with the inconsistencies reported in other studies [5].

Sensitivity to Hyperparameters In Appendix F, we investigate the sensitivity of our approach to hyperparameters, finding that it is less affected by factors such as the number of training images or different input prompts. Additionally, we demonstrate that our approach can leverage existing datasets to discover concept vectors.

5. Conclusion

In this study, we introduced a self-discovery approach to identify semantic concepts in the latent space of text-to-image diffusion models. Our research findings highlight that the generation of inappropriate content can be attributed to ethical-related concepts present in the internal semantic space of diffusion models. Leveraging these concept vectors, we enable responsible generation, including promoting equality among societal groups, eliminating inappropriate content, and enhancing text guidance for responsible prompts. Through extensive experiments, we have demonstrated the effectiveness and superiority of our proposed approach. Our work contributes to the understanding of internal representations in diffusion models and facilitates the generation of responsible content, maximizing the utility of high-quality text-to-image generation.

Acknowledgement This work is supported by the UKRI grant: Turing AI Fellowship EP/W002981/1, EPSRC/MURI grant: EP/N019474/1. We thank the Royal Academy of Engineering and FiveAI. This work is also founded by the German Federal Ministry of Education and Research and the Bavarian State Ministry for Science and the Arts.

References

[1] Manuel Brack, Felix Friedrich, Patrick Schramowski, and Kristian Kersting. Mitigating inappropriateness in image generation: Can there be value in reflecting the world's ugliness? *arXiv preprint arXiv:2305.18398*, 2023. 2

[2] 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. 2

[3] 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. 2

[4] Shreyansh Gandhi, Samrat Kokkula, Abon Chaudhuri, Alessandro Magnani, Theban Stanley, Behzad Ahmadi, Venkatesh Kandaswamy, Omer Ovenc, and Shie Mannor. Scalable detection of offensive and non-compliant content/logo in product images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2247–2256, 2020. 1

[5] Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*, 2023. 1, 2, 6, 7, 8

[6] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024. 1, 2, 5, 6

[7] Jindong Gu, Ahmad Beirami, Xuezhi Wang, Alex Beutel, Philip Torr, and Yao Qin. Towards robust prompts on vision-language models. *arXiv preprint arXiv:2304.08479*, 2023. 1

[8] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023. 2

[9] René Haas, Inbar Huberman-Spiegelglas, Rotem Mulayoff, and Tomer Michaeli. Discovering interpretable directions in the semantic latent space of diffusion models. *arXiv preprint arXiv:2303.11073*, 2023. 1, 2, 3, 4

[10] Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. *arXiv preprint arXiv:2305.10120*, 2023. 2

[11] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3, 12

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3, 12

[14] Jaeseok Jeong, Mingi Kwon, and Youngjung Uh. Training-free style transfer emerges from h-space in diffusion models. *arXiv preprint arXiv:2303.15403*, 2023. 3

[15] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022. 1

[16] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 2

[17] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702, 2023. 1, 2

[18] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 4

[19] Yipeng Leng, Qiangjuan Huang, Zhiyuan Wang, Yangyang Liu, and Haoyu Zhang. Diffusegae: Controllable and high-fidelity image manipulation from disentangled representation. *arXiv preprint arXiv:2307.05899*, 2023. 2

[20] Hang Li, Jindong Gu, Rajat Koner, Sahand Sharifzadeh, and Volker Tresp. Do dall-e and flamingo understand each other? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1999–2010, 2023. 1

[21] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 8

[22] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*, 2023. 2

[23] Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models. In *The Twelfth International Conference on Learning Representations*, 2023. 1

[24] Avery Ma, Amir-massoud Farahmand, Yangchen Pan, Philip Torr, and Jindong Gu. Improving adversarial transferability via model alignment. *arXiv preprint arXiv:2311.18495*, 2023. 1

[25] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 8, 15

[26] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 2

[27] Minheng Ni, Chenfei Wu, Xiaodong Wang, Shengming Yin, Lijuan Wang, Zicheng Liu, and Nan Duan. Ores: Open-vocabulary responsible visual synthesis. *arXiv preprint arXiv:2308.13785*, 2023. 2

[28] Zixuan Ni, Longhui Wei, Jiacheng Li, Siliang Tang, Yueting Zhuang, and Qi Tian. Degeneration-tuning: Using scrambled grid shield unwanted concepts from stable diffusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8900–8909, 2023. 2

[29] Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7053–7061, 2023. 2, 5

[30] Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. In *Advances in Neural Information Processing Systems*, 2023. 1, 2, 3, 4

[31] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 1

[32] Vinay Uday Prabhu and Abeba Birhane. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923*, 2020. 1

[33] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10619–10629, 2022. 2

[34] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[35] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1

[36] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022. 1

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 13

[38] Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1350–1361, 2022. 1, 6

[39] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023. 2, 6, 7

[40] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. *arXiv preprint arXiv:2309.11497*, 2023. 2

[41] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[42] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[43] Matthew Trager, Pramuditha Perera, Luca Zancato, Alessandro Achille, Parminder Bhatia, and Stefano Soatto. Linear spaces of meanings: Compositional structures in vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15395–15404, 2023. 2

[44] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 2

[45] Yingheng Wang, Yair Schiff, Aaron Gokaslan, Weishen Pan, Fei Wang, Christopher De Sa, and Volodymyr Kuleshov. Infodiffusion: Representation learning using information maximizing diffusion models. *arXiv preprint arXiv:2306.08757*, 2023. 2

[46] Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. Uncovering the disentanglement capability in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2023. 2

[47] Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. Itigen: Inclusive text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3969–3980, 2023. 2

[48] Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591*, 2023. 2

[49] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1

[50] Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. *arXiv preprint arXiv:2310.11868*, 2023. 2

[51] Zijian Zhang, Zhou Zhao, and Zhijie Lin. Unsupervised representation learning from pre-trained diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 35:22117–22130, 2022. 2

[52] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018. 5, 6