Unified Concept Editing in Diffusion Models

Rohit Gandikota[1] Hadas Orgad[2] Yonatan Belinkov[2] Joanna Materzyńska[3] David Bau[1]
[1]Northeastern University [2]Technion [3]Massachusetts Institute of Technology

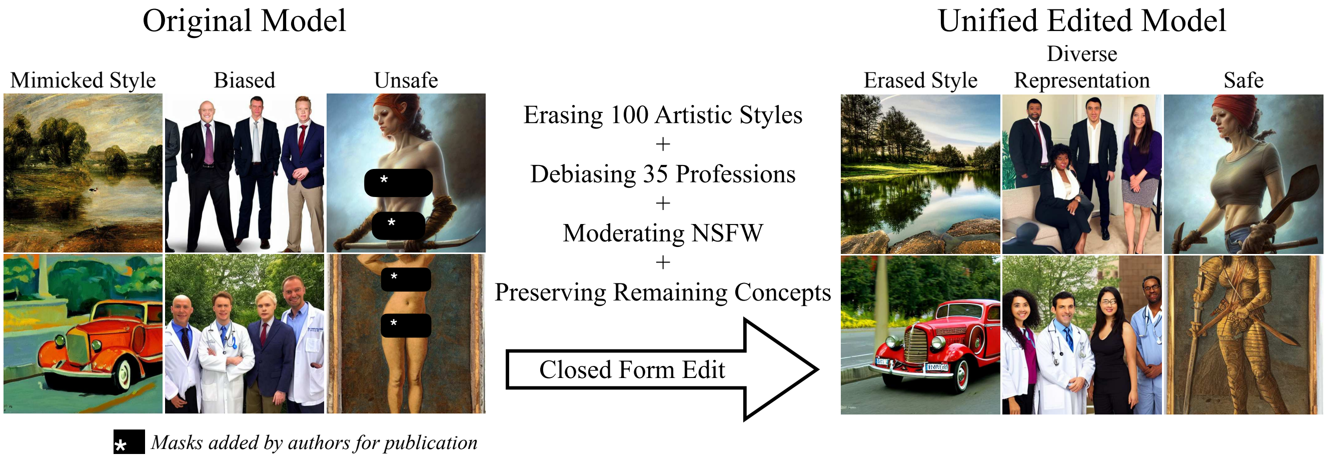

Figure 1. Our method enables unified and efficient editing of multiple concepts in text-to-image models through closed-form modifications to attention weights. We present applications to debias, erase, and moderate concepts at scale. Debiasing professions leads the edited model to generate fairer gender and race ratios. Erasing an artistic style removes characteristics associated with a particular creator. Moderating the model reduces the likelihood of generating inappropriate images.

Abstract

Text-to-image models suffer from various safety issues that may limit their suitability for deployment. Previous methods have separately addressed individual issues of bias, copyright, and offensive content in text-to-image models. However, in the real world, all of these issues appear simultaneously in the same model. We present a method that tackles all issues with a single approach. Our method, Unified Concept Editing (UCE), edits the model without training using a closed-form solution, and scales seamlessly to concurrent edits on text-conditional diffusion models.

We present scalable simultaneous debiasing, style erasure, and content moderation by editing text-to-image projections, and perform extensive experiments demonstrating improved efficacy and scalability over prior work. Our code is available at unified.baulab.info.

1. Introduction

Text-to-image diffusion models have ushered in a set of complex societal challenges. Generative image models jeopardize artists by cloning their styles [1]; they reinforce biases by amplifying stereotypes [24, 43]; and they facilitate the creation of offensive images [18]. While several methods have been proposed to mitigate such issues individually [8, 13, 21, 35, 45], real-world deployments of generative image models manifest all these problems concurrently. A natural first step for exercising safety in generative models is the careful curation of training data to exclude any content that should not be replicated [33]. However, training a large model is expensive, and the impact of data curation on a model may be counterintuitive and unpredictable. For example, removing undesired content can expose other undesired content [6]; removing toxic content can introduce new biases [10]; and reducing offensive content can result in incomplete removal [28]; these examples highlight the limitations of relying solely on data curation.

In this paper, we introduce a unified model-editing approach capable of addressing the different safety issues with a single formulation. Our method, called *Unified Concept Editing* (UCE), offers a fast and practical way to control model behavior post-training, filling the gaps where data curation might fall short. UCE is a closed-form parameter-editing method that enables the application of hundreds of editorial modifications within a single text-to-image synthesis model while preserving the generative quality of the model for unedited concepts.

The UCE method builds upon previous model editing work, generalizing the TIME [30] and MEMIT [26] methods. Unlike previous diffusion model editing methods such as TIME, UCE is designed to enable many simultaneous edits to be applied at once. These edits can include actions such as erasing, moderating, or debiasing a concept—tasks that have been traditionally treated as distinct issues with separate solutions. UCE goes beyond MEMIT in several ways: it edits text-to-image models rather than language models; and it also allows the editor to explicitly specify the distribution of concepts that should not be modified. Finally, UCE also introduces a new, scalable debiasing approach. We compare UCE with a range of prior model-editing methods and find that it demonstrates superior performance, outperforming other methods by a wide margin. UCE exhibits superior performance both in single edits in each category of editing, as well as in the ability to scale to many edits at once while minimizing interference with unedited concepts.

2. Related Work

While text-to-image diffusion models are becoming increasingly popular in commercial art and graphic design, they tend to suffer from various issues, which have previously been addressed separately.

Copyright issues. Recent lawsuits [1, 37] have contended that models like Stable Diffusion infringe on many artistic styles, and researchers have found that the models can memorize some copyrighted training data nearly verbatim [5, 41]. To reduce such memorization, previous work proposes randomizing and augmenting training image captions [42], while other work has explored a technique called image cloaking that allows artists to protect their content from being imitated by large generative models by adding specially crafted adversarial perturbations to images before publishing them online [34, 38]; both these approaches require thorough preparation of the training corpus. Another approach adjusts a model after training is complete, deleting an undesired concept by modifying model weights [13, 15, 20, 21, 46]. Our method adopts that concept-erasure approach, and we benchmark against the previous state-of-the-art. Our method differs from previous concept erasure methods because it is a closed-form edit that removes many concepts at once.

Offensive content. Diffusion models also sometimes generate inappropriate images, such as nude and violent images. Various methods have been proposed to filter out inappropriate images from the training data or at inference time [12, 27]; for example the Stable Diffusion implementation includes a "not safe for work" safety checker that returns a black image when an unsafe image is detected [3, 22, 32], and other work has addressed the issue in through image editing at inference time [35]. In cases where open-source code and model weights are openly available, such post-production filters can be easily disabled [39]. A more difficult-to-circumvent approach removes the knowledge of unwanted concepts from the model weights; previous methods taking that approach have proposed attention re-steering through fine-tuning [46], fine-tuning the attention weights [13] and continual learning [15]. While previous methods all fine-tune the model, we propose a fast and efficient method to erase offensive concepts using a closed-form edit.

Social biases. Diffusion image generation models have been found to be prone to social and cultural biases [7, 24, 43], even exaggerating and amplifying societal stereotypes beyond simple imbalances in the training data [4, 11], although quantifying amplification can be subtle [36]. Previous work has tackled this issue by modifying model parameters after training, by projecting out biased directions in the text embedding [8], or by performing algebraic manipulation of the representations [45]. One previous work, which inspires our current method, applies a direct closed-form model editing method [30]. The previous works have found that debiasing multiple concepts simultaneously is challenging, because debiasing one concept affects others, even in the presence of regularization methods. Our method overcomes that limitation with a new debiasing procedure that eliminates the mutual effect between concepts.

Model editing. Model editing has recently emerged as an approach to control a model's behavior without training. In model editing, a subset of the model's weights is modified by locating the knowledge in the model and targeting it. Closed-form solutions for editing knowlege in generative text models have been proposed in [25, 26], while [2, 30] have edited knowledge in text-to-image diffusion models by targeting either the cross-attention layers or the text-encoder layers. Our method adopts and generalizes these approaches to enable removal and debiasing of many concepts simultaneously in text-to-image models.

3. Background

Diffusion models are generative models that can approximate distributions through a gradual denoising process [16, 40]. Starting from Gaussian noise, the model iteratively denoises over T time steps to form a final image. At each intermediate step t, the model predicts noise ϵ_t that is added to the original image, with x_T as initial noise and x_0 as the final output. By learning the parameters of the denoising process, the trained model can generate novel images from noise. This denoising is modeled as a Markov transition probability.

$$p_\theta(x_{T:0}) = p(x_T) \prod_{t=T}^{1} p_\theta(x_{t-1}|x_t) \tag{1}$$

Text-to-image latent diffusion models operate on low-dimensional embedding that is modeled with a U-Net generation network. The text conditioning is fed to the network via text embedding, extracted from a language model, in the cross-attention layers. Specifically, the attention modules within diffusion models follow the QKV (Query-Key-Value) [44] structure, where queries originate from the image space, while keys and values are derived from the text embeddings. Our focus centers on the linear layers W_k and W_v, responsible for projecting text embeddings.

For a given text embedding c_i, the keys and values are generated by $k_i = W_k c_i$ and $v_i = W_v c_i$ respectively. The keys are then multiplied by the query q_i that represents the visual features of the current intermediate image. This produces an attention map that aligns relevant text and image regions:

$$\mathcal{A} \propto \text{softmax}(q_i k_i^T) \tag{2}$$

The attention map indicates the relevance between each text token and visual feature. Using this alignment, the cross-attention output is then computed by attending over the value vector V with the normalized attention weights.

$$\mathcal{O} = \mathcal{A} v_i \tag{3}$$

The cross-attention is the mechanism that links the text and image information and responsible for assigning visual meaning to text tokens. The output of Equation 3 is then propagated through the remaining layers of the diffusion U-Net.

TIME [30] edits implicit assumptions in pre-trained diffusion models by updating the cross-attention layers. Implicit assumptions can be any visual features that a model assumes about objects in an under-specified prompt, such as the color of roses which is usually red, or the gender of a doctor which is usually male. To edit these assumptions, the method requires a "source" under-specified prompt where the model makes an assumption (e.g. "a pack of roses") and a "destination" prompt specifying the desired attribute (e.g. "a pack of blue roses"). TIME updates the projection matrices W_k and W_v, to bring the source prompt embedding closer to the destination embedding. This aligns the textual concepts such that the model no longer makes the implicit assumption.

Let c_i be the source embedding, derived from the tokens of the source prompt, and c_i* be the corresponding destination embeddings, taken from the embeddings of the corresponding tokens in destination prompt. The values of the destination prompts are calculated as $v_i* = W^{\text{old}} c_i*$. New projection matrices W are then optimized to minimize the objective function (a similar equation for the key projection matrices can be derived):

$$\min_W \sum_{i=0}^{m} ||Wc_i - \underbrace{v_i^*}_{W^{\text{old}} c_i^*} ||_2^2 + \lambda ||W - W^{\text{old}}||_F^2 \tag{4}$$

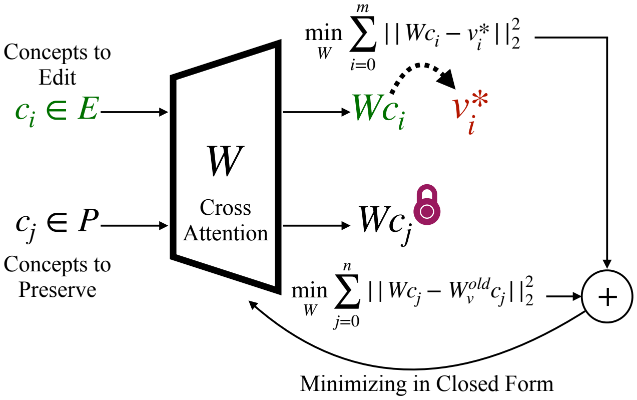
$\min_W \sum_{i=0}^{m} || Wc_i - v_i^* ||_2^2$
Concepts to Edit
$c_i \in E$
W
Cross Attention
Wc_i
v_i^*
$c_j \in P$
Wc_j
Concepts to Preserve
$\min_W \sum_{j=0}^{n} || Wc_j - W_v^{old} c_j ||_2^2$
+
Minimizing in Closed Form

Figure 2. Closed-form editing of cross-attention weights enables concept manipulation in diffusion models. Our method modifies the attention weights to induce targeted changes to the keys and values corresponding to specific text embeddings for a set of edited concepts $c_i \in E$ while minimizing changes to a set of preserved concepts $c_j \in P$. That dual objective allows debiasing, erasing, or moderating concepts while preserving unrelated ones. The same editing function is applied in all cases, but the target keys and values are set differently per application. As a closed-form edit, modifying attention weights given the new keys and values mappings takes less than 1 minute. That enables efficient simultaneous editing of multiple concepts.

where λ is a regularization hyper-parameter. [30] proved that the loss function has a closed-form global minimum solution, which allows efficient editing of text-to-image models.

$$W = \left(\sum_{i=0}^{m} v_i^* c_i^T + \lambda W^{\text{old}} \right) \left(\sum_{i=0}^{m} c_i c_i^T + \lambda \mathbb{I} \right)^{-1} \tag{5}$$

The first term in the inverse matrix, $\sum_{i=0}^{m} c_i c_i^T$, is the covariance of the concept text embeddings being edited. As discussed in the appendix, we interpret the second term, an identity matrix, as matching the covariance of the large encyclopedia of concept embeddings in the diffusion model's vocabulary, inspired by MEMIT [26].

While TIME formulation is effective, it risks interference with surrounding concepts when editing a particular concept. For example, editing doctors to be female might also affect teachers to be female. TIME has a regularization term that prevents the edited matrix from changing too radically. However, it is a general term and thus affects all vector representations equally. In this work, we present an alternative preservation term that allows targeted editing of the parameters of the pretrained generative model while maintaining its core capabilities.

4. Method

We introduce a general model editing methodology applicable to any linear projection layer. Given a pretrained layer W^{old}, our goal, as shown in Figure 2, is to find new edited weights W that edit a set of concepts in set E while preseving a set of concepts in set P. Specifically, we wish to find weights so that the output for each of the inputs $c_i \in E$ maps to target values $v_i^* = W_v^{\text{old}} c_{i^*}$ instead of the original $W^{\text{old}} c_i$, while preserving outputs corresponding to the inputs $c_j \in P$ as $W^{\text{old}} c_j$. A formal objective function can be constructed as:

$$\min_W \sum_{c_i \in E} ||W c_i - v_i^*||_2^2 + \sum_{c_j \in P} ||W c_j - W^{\text{old}} c_j||_2^2 \tag{6}$$

As derived in the Appendix, the objective function in Equation 6 has a closed-form solution for the updated weights:

$$W = \left(\sum_{c_i \in E} v_i^* c_i^T + \sum_{c_j \in P} W^{\text{old}} c_j c_j^T \right) \left(\sum_{c_i \in E} c_i c_i^T + \sum_{c_j \in P} c_j c_j^T \right)^{-1} \tag{7}$$

This formulation generalizes both the TIME [30] and MEMIT [26] editing methods. When only canonical directions of the inputs are used as preservation terms c_j, our method reduces to TIME. Solving for the weight update ΔW instead of directly solving for W, our method reduces to MEMIT closed-form update. We discuss in detail how our approach provides a unified generalization that encompasses prior editing techniques as special case in the Appendix.

We edit the linear cross-attention projections (W_k and W_v) to perform various concept edits with different goals: erasure, moderation, and debiasing. Our method requires the m text embeddings c_i derived from text descriptions of the concepts to edit and their corresponding modified target outputs v_i^*. The target outputs are defined differently based on the edit type, through the destination concepts c_i^* as described below. We also preserve n surrounding concepts using their descriptions c_j. For concepts with multiple tokens, we align the last token of c_i to the last token of v_i^* and make the edit.

Erasing To erase a concept c_i, we want to prevent the model from generating it. If the concept is abstract like an artistic style (e.g. "Kelly Mckernan"), this can be accomplished by modifying the weights so the target output v_i aligns with a different concept c_* (e.g. "art"):

$$v_i^* \leftarrow W^{\text{old}} c_* \tag{8}$$

This updates the weights such that the output no longer reflects concept c_i, effectively erasing that concept from the model's generations and eliminating generations of the undesired characteristics.

Debiasing To debias a concept c_i (e.g. "doctor") across attributes $a_1, a_2, ..., a_p$ (e.g. "white", "asian", "black", ..), we want the model to generate the concept with evenly distributed attributes. This is achieved by adjusting the magnitude of v_i along the directions of $v_{a_1}, v_{a_2}, ..., v_{a_p}$, where $v_{a_i} = W^{\text{old}} a_i$ corresponds to the attribute text prompts:

$$v_i^* \leftarrow W^{\text{old}} \left[c_i + \alpha_1 a_1 + \alpha_2 a_2 + ... + \alpha_p a_p \right] \tag{9}$$

The constants α_i are chosen such that the diffusion model generates the concept with any desired probability for each attribute. This enables our method to debias multiple attributes simultaneously, unlike previous approaches such as TIME and concept ablation that can debias across dual attributes only. We provide the detailed algorithm in Alg 1.

Algorithm 1 Debiasing Concepts in Diffusion Models

```
Input: Diffusion $M$ with cross attentions $W_k, W_v$
Input: Edit list $E$, preserve list $P$
Input: Attributes $A$ (list of strings of size $p$)
Input: Learning step $\eta$, desired ratios $R_{des}$
while True do
    $R_{curr} \leftarrow$ GET_RATIOS($M, E, A$)
    for $i, c_i$ in enumerate($E$) do
        if $max(|R_{curr}[i] - R_{des}[i]|) < 0.05$ then
            $P$.append($c_i$)   ▷ add to preserve post debias
            $E$.remove($c_i$)   ▷ remove from edit list
            continue
        end if
        $\alpha \leftarrow \eta(R_{curr}[i] - R_{des}[i])$   ▷ $\alpha \in \mathcal{R}^p$
        $v_i^* \leftarrow W_v c_i + \alpha \cdot A$
        $k_i^* \leftarrow W_k c_i + \alpha \cdot A$
    end for
    if $E$ is empty then
        break   ▷ All concepts debiased
    end if
    $W_v$ = UCE($E$, $\{v_i^*\}$, $P$, $W_v$)   ▷ UCE is Eq.7
    $W_k$ = UCE($E$, $\{k_i^*\}$, $P$, $W_k$)
end while
return $M$   ▷ Debiased Model
```

Moderation To moderate concept c_i (e.g. "nudity"), we perform an edit where the target output v_i^* aligns with an unconditional prompt c_0 (e.g. " "):

$$v_i^* \leftarrow W^{\text{old}} c_0 \tag{10}$$

This replaces the output for c_i with a more generic, unconditional output c_0, moderating the model's response by reducing extreme attributes of that concept.

5. Experiments

5.1. Erasing

Our erasing technique directly modifies the key–value mappings in the model to associate keys with different concepts rather than the undesired ones. We use our method to erase artistic styles from the model's weights. Our technique allows preserving certain artists while removing others. We found this enables substantially less interference on a holdout set of artists that were neither erased nor explicitly preserved. We compare our artistic erasure method to recent approaches including ESD-x [13], Concept Ablation [21], and SDD [20] which use cross-attention fine-tuning for controllable image editing. In a second set of experiments, we test object erasure (e.g., erasing the concept of garbage trucks). In this set of experiments, we did not use any explicit preservation objectives, in order to test implicit interference. For object erasure, we primarily compare to ESD-u [13], which freezes all parameters except cross-attentions during fine-tuning, enabling more global erasures.

5.1.1 Artist erasure

Our method can successfully erase multiple concepts while preserving the model's knowledge. We use the text embeddings of the artist names as our concepts c_i to erase and a set of artists to preserve c_j. As shown in Figure 3, we are able to consistently erase multiple artistic styles, while other methods maintain a lot of characteristics of the artistic styles and impair the model's capabilities as the number of erased concepts increases. We found ESD and SDD tend to damage the model more when erased sequentially (at 1000 iterations per concept), so we limited those techniques to random sampling-based erasure for a fixed 1000 iterations.[1]

Our method also demonstrates reduced interference with neighboring, non-erased concepts compared to other techniques. As shown in Figure 4, erasing with our approach has less impact on concepts that were not targeted for removal. The top plot shows the LPIPS [47] difference between the original SD and edited models, indicating our method results in the smallest modifies the unrelated concepts the least. The bottom plot shows the CLIP score [31] on COCO-30k prompts [23], where our method maintains better text-to-image alignment after editing, as evidenced by the higher CLIP score. Together, these results highlight an important advantage of our erasing approach – the ability to remove targeted concepts with minimal disruption to other areas of knowledge in the model.

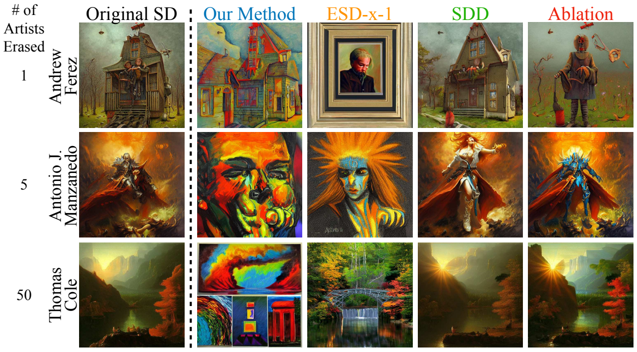

Figure 3. Our method and ESD-x show strong erasing capabilities. SDD and Ablation[2]start to dilute their erasing capabilities as the number of concepts being erased are increased.

# Concepts	CLIP ↑	LPIPS↓	FID↓
1	31.35	0.05	14.37
5	31.25	0.08	14.30
10	31.48	0.13	15.56
50	31.22	0.22	15.10
100	30.08	0.30	15.09
500	21.06	0.44	72.40
1000	16.79	0.47	85.48
Original SD	31.32	-	14.49

Table 1. Our method can erase upto 100 concepts while performing similar to pre-trained SD on COCO-30k dataset. The image fidelity is consistent with original SD till 100 erasures. With LPIPS, we find that the model at 100 erasures has a slightly different performance for a given seed and prompt, but as the CLIP score shows, the alignment of the model is still intact.

Diffusion models were shown to mimic more than 1800 artistic styles [19]. We analyzed the capabilities of our method to erase multiple concepts by erasing n artists while preserving the remaining $1000 - n$. As shown in Table 1, our method can erase up to 100 artists simultaneously before damaging image fidelity and CLIP scores. After 50 erasures, the model's output for a given prompt and seed begins to change , as indicated by the LPIPS score, but remains aligned overall as evidenced by the CLIP score. The importance of our preservation strategy to these results is shown in the Appendix, where no preservation reduces back to the TIME formulation.

5.1.2 Erasing Objects

To demonstrate the capability of our method to erase objects from the diffusion model's learned concepts, with potential applications for removing harmful symbols and content, we conducted experiments erasing Imagenette [17] classes, a subset of Imagenet classes [9]. For each erased object, we utilized the text embedding (e.g. "French Horn") as c_i, without additional preservation concepts c_j. We generated 500 images per class and evaluated top-1 classification accuracy

[1] The authors of SDD note potential overfitting of their method when erasing multiple concepts. To mitigate this, we limited SDD to 700 iterations for multi-concept erasure. For Ablation, the authors suggest 100 iterations per concept, however we found the model deteriorates after 1000 total iterations. Therefore, we restricted Ablation to 1000 iterations total when erasing multiple concepts.

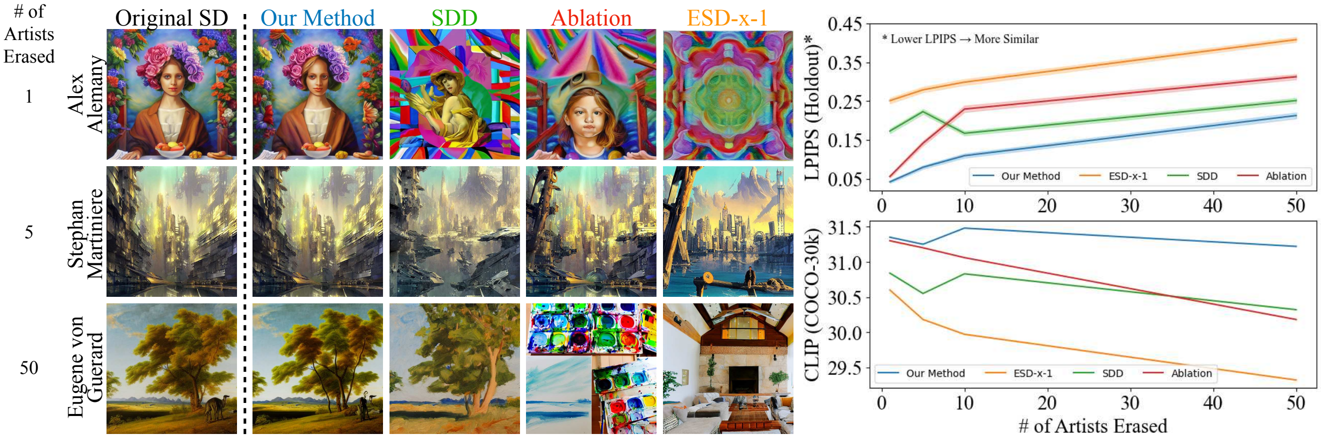

Figure 4. Our method preserves the remaining knowledge of the model better after the edit. The figure shows images generated from different editing methods, for the same prompts and seeds, across a variety of **artists that are not erased**. Our method exhibits lower LPIPS, indicating less change to unerased concepts during model editing. Similarly for COCO, we find that our method has better CLIP scores across all the scales. This demonstrates that our method has significantly reduced interference compared to other fine-tuning approaches when editing.

Class name	**Accuracy of Erased Class ↓**			**Accuracy of Other Classes ↑**		
	SD	**Ours**	**ESD-u**	**SD**	**Ours**	**ESD-u**
Cassette Player	15.6	0.0	0.60	85.1	90.3	64.5
Chain Saw	66.0	0.0	6.0	79.6	76.1	68.2
Church	73.8	8.4	54.2	78.7	80.2	71.6
Gas Pump	75.4	0.0	8.6	78.5	80.7	66.5
Tench	78.4	0.0	9.6	78.2	79.3	66.6
Garbage Truck	85.4	14.8	10.4	77.4	78.7	51.5
English Springer	92.5	0.2	6.2	76.6	78.9	62.6
Golf Ball	97.4	0.8	5.8	76.1	79.0	65.6
Parachute	98.0	1.4	23.8	76.0	77.4	65.4
French Horn	99.6	0.0	0.4	75.8	77.0	49.4
Average	78.2	**2.6**	12.6	78.2	**79.8**	63.2

Table 2. Our method can erase objects from diffusion models effectively without impacting the accuracy for other object classes even when they are not explicitly preserved. Compared to ESD-u, we demonstrate improved erasure of the targeted class alongside higher preservation of unrelated classes in the generated images on Imagenette classes.

using a pretrained ResNet-50 [14], comparing to ESD-u in Table 2. Objects were erased individually to analyze interference versus ESD on non-erased classes. Without explicit preservation, our approach exhibited superior erasure capability while minimizing interference on non-targeted classes. Further erasure analysis is provided in the Appendix. Erasing all 10 Imagenette classes together reduced image generation accuracy to just 4.0% and COCO-CLIP score to 31.02 (original SD is 31.32), quantitatively showing effective single and multi-object removal while limiting interference.

5.2. Debiasing

Stable Diffusion exhibits gender and racial bias when generating images for profession names (e.g. CEO), producing only 6% female figures for "CEO" prompt. We debias profession concepts via Alg. 1, using profession text embeddings c_i and attribute embeddings A (e.g. "male", "female"). To prevent over/under-debiasing, we set per-attribute regularization constants α_i in Eq. 9. As debiasing one concept can affect others [30], we use an iterative approach. We maintain edit and freeze concept lists, fixing debiased concepts while editing new ones. With multiple concepts debiased in parallel, α_i values are found by generating validation samples during training and adjusting constants based on the current model's generated ratio (classified by CLIP). Once a concept is sufficiently debiased, we add it to a preservation list, bypassing validation and keeping it fixed when debiasing others. This iterative α_i tuning enables efficient debiasing by avoiding unnecessarily repeated editing of already debiased concepts. Setting equal α_i for all concepts risks over-debiasing some while under-debiasing others. Our iterative validation determines optimal per-concept constants.

5.2.1 Gender bias

Prior methods for debiasing generative models like TIME [30], Concept Algebra [45], and Debiasing-VL [8] have focused on mitigating biases between two discrete attributes. While we acknowledge that a binary perspective of gender excludes non-binary groups, for a fair comparison to such dual-attribute techniques, we evaluate our method by reducing occupational gender biases in diffusion models. We recognize that editing for visual features of non-binary genders risks introducing other unwanted stereotypical behavior.

Figure 5 provides qualitative results demonstrating increased diversity in generated images for professions with strong initial gender biases after applying our proposed debiasing technique. For quantitative evaluation, we synthesize

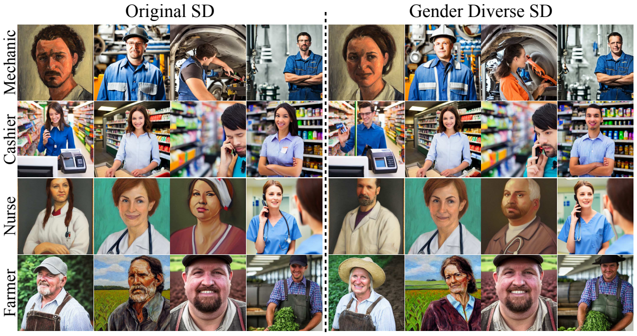

Figure 5. Our method improves the gender representation of professions in the stable diffusion generated images. We find that the images precisely change the gender while keeping the rest of the scene intact.

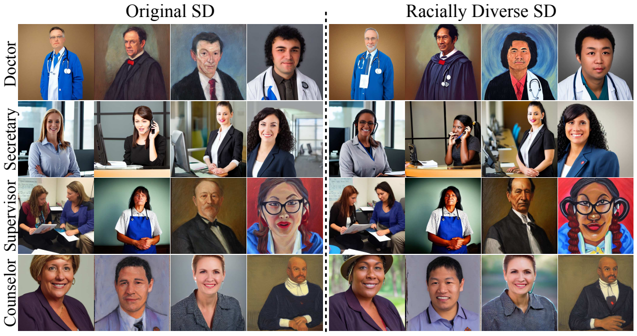

Figure 6. Our method improves the racial diversity of professions in the pre-trained stable diffusion. We show images from the original SD and the corresponding images from the edited model for the same prompts and seeds for comparison. We find that our edited model has a better race representation.

250 images per profession and utilize CLIP classifications to calculate the deviation $\Delta = \frac{|p_{\text{desired}} - p_{\text{actual}}|}{p_{\text{desired}}}$ between the achieved and desired (50-50) gender ratios, where $\Delta = 0$ indicates perfect debiasing. As shown in Table 3, our method achieves gender distributions closest to the balanced 50-50 ratio compared to pretrained and baseline models. The original formulation of TIME [30] exhibits interference between debiased concepts, resulting in worse performance. We find that even when applying TIME with our proposed preservation term, it still underperforms compared to our approach. Through both qualitative and quantitative results, we demonstrate that our method enables robust targeted debiasing of generative models.

5.2.2 Racial bias

A key advantage of our approach over prior debiasing techniques is the ability to concurrently mitigate biases related to multiple attributes. To demonstrate this capability, we conduct experiments to improve racial diversity in professions generated by Stable Diffusion. Specifically, we target major racial categories as defined by U.S. Office of Management and Budget (OMB) standards [29]: White, Black, American Indian, Native American, and Asian. Accurately classifying race from images is an intricate task, problematic even for sophisticated models like CLIP and humans. We, therefore, take a qualitative analysis approach rather than attempting error-prone quantitative race categorization. As depicted in Figure 6, our method significantly enhances the representation of these racial groups among generated professional images. This highlights our technique's strength in reducing multifaceted biases in diffusion models, a key advantage over existing binary-attribute debiasing methods.

5.3. Moderation

We quantitatively evaluate our proposed method for moderating sensitive concepts, comparing it against recent state-of-the-art techniques ESD-u and ESD-x [13] on the task of erasing single concepts like "nudity". For all the models, 4703 images are generated using the prompts from the I2P benchmark introduced in [35]. In Figure 7 we analyze the nudity moderation using NudeNet classifier [3]. We find that our method demonstrates comparable nudity erasure performance to ESD-X since both techniques edit cross-attentions of the diffusion model. ESD-u as expected has a more aggressive erasure effect given it finetunes the entire model except cross attentions. However, Table 4 highlights that our approach induces substantially lower distortion to model generations than ESD-u and ESD-x, with significantly reduced LPIPS [47] score from the original SD generations. This indicates our method better preserves image quality while moderating sensitive concepts. Additionally, the CLIP score indicates that our technique maintains better text-image alignment post editing.

We further demonstrate efficacy in erasing multiple sensitive concepts from I2P [3]. Again, our approach shows improved multi-concept moderation capability compared to ESD-u (Figure 7). We provide a detailed analysis of moderating diverse sensitive concepts in the Appendix.

5.4. Unified Editing

Our formulation enables simultaneous style erasure, profession debiasing, and nudity moderation. The edit vector v^* design differs for each edit, but the model update is unified. Empirically, the jointly finetuned model demonstrates effectiveness on par with individually trained models: similar to Table 3, a gender ratio deviation of 0.27 versus 0.22 for the gender-debiasing model and 0.67 for the original Stable Diffusion. The unified model also shows a 58% nudity reduction compared to 49% for nudity erasure and 64% for ESD-u, shown in Figure 7.

[3] including hate, harassment, violence, suffering, humiliation, harm, suicide, sexual, nudity, bodily fluids, blood, obscene gestures, illegal activity, drug use, theft, vandalism, weapons, child abuse, brutality, cruelty

Profession	Original-SD	Concept Algebra	Debias-VL	TIME	TIME + Preserve	Ours
Librarian	0.86 ± 0.06	0.66 ± 0.07	0.34 ± 0.06	0.26 ± 0.05	0.35 ± 0.01	**0.07** ± **0.07**
Teacher	0.42 ± 0.01	0.46 ± 0.00	0.11 ± 0.05	0.34 ± 0.06	0.07 ± 0.06	**0.06** ± **0.02**
Sheriff	0.99 ± 0.01	0.38 ± 0.22	0.82 ± 0.08	0.22 ± 0.05	0.10 ± 0.05	**0.10** ± **0.03**
Analyst	0.58 ± 0.12	0.24 ± 0.18	0.71 ± 0.02	0.52 ± 0.03	**0.13** ± **0.05**	0.20 ± 0.07
Doctor	0.78 ± 0.04	0.40 ± 0.02	0.50 ± 0.04	0.58 ± 0.03	0.41 ± 0.08	**0.20** ± **0.02**
WinoBias [48]	0.67 ± 0.01	0.43 ± 0.01	0.55 ± 0.01	0.44 ± 0.00	0.31 ± 0.00	**0.22** ± **0.00**

Table 3. Debiasing performance on 5 randomly-picked professions and an average on all 35 Winobias [48] professions. The presented metric Δ measures the percentage deviation from desired ratios ($\Delta = 0$ indicates complete debiasing). Our method has a consistent debiasing performance compared to previous inference and model editing methods by showing the least average deviation from the desired distribution.

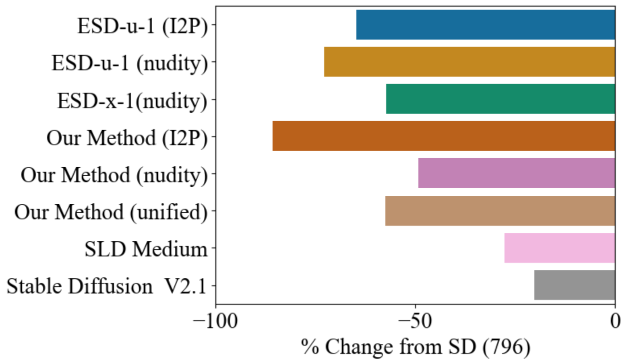

Figure 7. A percentage reduction in nudity-classified samples on I2P prompts compared to SD. Our method erases nudity content from pre-trained SD and has the advantage of erasing multiple concepts in I2P prompts. "`Nudity`" erased model performs very similar to `ESD-x-1` as both the methods edit only cross attentions. However, as noted in Table 4, we find that our method results in a finer edit and has better alignment with COCO.

Method	FID-Real ↓	FID-SD ↓	CLIP ↑	LPIPS ↓
REAL	-	14.49	30.41	-
SD	14.49	-	31.32	-
ESD-u-1	**14.16**	3.73	30.45	0.23
ESD-x-1	14.45	2.33	30.81	0.18
Ours	14.84	**1.82**	**31.26**	**0.12**

Table 4. Our method performs comparably to the pre-trained SD on COCO. The image fidelity performance compared to SD (FID-SD) and LPIPS matches closely with our method. FID with real COCO images (FID-real) is very similar to SD. Our method also has the closest CLIP score to the original SD compared to other methods.

6. Limitations

When debiasing across multiple attributes, we find interdependencies that exhibit compounding biases. For example, generating images of "a black person" has near equal gender ratios (48% male out of 100 images), while "a native american person" displays strong male bias (96% male of 100). Debiasing in isolation can thus perpetuate biases along other dimensions. This highlights the need for joint attribute consideration to mitigate propagated biases. We also find word-level biases in prompts that compose unfavorably. Non-gendered phrases like "successful person" become predominantly male (88% of 100) versus the gender-balanced "person" (50% male of 100), illustrating how subtle cues carry biases. Such compositional effects pose challenges, as each word element contributes biases needing mitigation.

For artistic style erasure, removing over 500 artists degrades general image generation, even with preservation terms (Table 1). That suggests a critical mass of artists is needed to maintain generative capabilities. Excessive erasure impairs the core visual priors learned during pretraining.

7. Conclusion

We have presented a unified algorithm for precisely editing diffusion models to allow designers to make them more responsible and beneficial for society. Our approach enables targeted debiasing, erasure of potentially copyrighted content, and moderation of offensive concepts, using only text descriptions. Our measurements suggest that our method offers three key benefits over prior methods. First, it can mitigate multifaceted gender, racial, and other biases simultaneously while preserving model capabilities. Second, it is scalable, modifying hundreds of concepts in one pass without expensive retraining. Third, extensive experiments demonstrate superior performance on real-world use cases. Together, our findings suggest that UCE is significant step towards democratizing access to ethical and socially-responsible generative models. The ability to seamlessly unify debiasing, erasure, and moderation will be an important tool for building AI that benefits our diverse global society.

Acknowledgments

Thanks to Antonio Torralba for valuable advice, discussions, and support. RG and DB are supported by Open Philanthropy. HO and YB are supported by the Israel Science Foundation (grant No. 448/20), Open Philanthropy, and an Azrieli Foundation Early Career Faculty Fellowship. The Center for AI Safety made computing resources available for this research.

References

[1] Sarah Andersen, Kelly McKernan, and Karla Ortiz. et al v. Stability AI Ltd. et al. Case No. 3:2023cv00201. US District Court for the Northern District of California., Jan 2023. 1, 2

[2] Dana Arad, Hadas Orgad, and Yonatan Belinkov. Refact: Updating text-to-image models by editing the text encoder. *arXiv preprint arXiv:2306.00738*, 2023. 2

[3] Praneeth Bedapudi. NudeNet: Neural nets for nudity detection and censoring, 2022. 2, 7

[4] 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 *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1493–1504, New York, NY, USA, 2023. Association for Computing Machinery. 2

[5] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023. 2

[6] Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, and Florian Tramer. The privacy onion effect: Memorization is relative. *Advances in Neural Information Processing Systems*, 35:13263–13276, 2022. 1

[7] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative models. *arXiv preprint arXiv:2202.04053*, 2022. 2

[8] 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. 1, 2, 6

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[10] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, 2018. 1

[11] Kathleen C. Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. A friendly face: Do text-to-image systems rely on stereotypes when the input is under-specified? In *The AAAI-23 Workshop on Creative AI Across Modalities*, 2023. 2

[12] 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 *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2236–2245, 2020. 2

[13] 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, 5, 7

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

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

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[17] Jeremy Howard and Sylvain Gugger. Fastai: A layered api for deep learning. *Information*, 11(2):108, 2020. 5

[18] Tatum Hunter. Ai porn is easy to make now. for women, that's a nightmare., 2 2023. 1

[19] Surea I, Proxima Centauri B, Erratica, and Stephen Young. Image synthesis style studies, 7 2022. 5

[20] Sanghyun Kim, Seohyeon Jung, Balhae Kim, Moonseok Choi, Jinwoo Shin, and Juho Lee. Towards safe self-distillation of internet-scale text-to-image diffusion models. *arXiv preprint arXiv:2307.05977*, 2023. 2, 5

[21] 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 2023 IEEE International Conference on Computer Vision*, 2023. 1, 2, 5

[22] Gant Laborde. NSFW detection machine learning model, 2022. 2

[23] 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. 5

[24] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023. 1, 2

[25] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022. 2

[26] Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3, 4

[27] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[28] Ryan O'Connor. Stable diffusion 1 vs 2 - what you need to know, 2022. 1

[29] Office of Management and Budget. Office of management and budget (omb) standards. U.S. Department of Health and Human Services, Jul 2022. 7

[30] Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*, 2023. 2, 3, 4, 6, 7

[31] 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. 5

[32] Javier Rando, Daniel Paleka, David Lindner, Lennard Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022. 2

[33] Robin Rombach. Stable diffusion 2.0 release, Nov 2022. 1

[34] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. *arXiv preprint arXiv:2302.06588*, 2023. 2

[35] 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. 1, 2, 7

[36] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. 2023. 2

[37] Riddhi Setty. Ai art generators hit with copyright suit over artists' images, 1 2023. 2

[38] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by text-to-image models. *arXiv preprint arXiv:2302.04222*, 2023. 2

[39] SmithMano. Tutorial: How to remove the safety filter in 5 seconds, 8 2022. 2

[40] 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. 2

[41] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023. 2

[42] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *arXiv preprint arXiv:2305.20086*, 2023. 2

[43] Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. The biased artist: Exploiting cultural biases via homoglyphs in text-guided image generation models. *arXiv preprint arXiv:2209.08891*, 2022. 1, 2

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[45] Zihao Wang, Lin Gui, Jeffrey Negrea, and Victor Veitch. Concept algebra for text-controlled vision models. *arXiv preprint arXiv:2302.03693*, 2023. 1, 2, 6

[46] 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

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5, 7

[48] 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. 8