Editing Implicit Assumptions in Text-to-Image Diffusion Models

Hadas Orgad* Bahjat Kawar* Yonatan Belinkov†
Computer Science Faculty, Technion, Israel
{orgad.hadas@cs., bahjat.kawar@cs., belinkov@}technion.ac.il

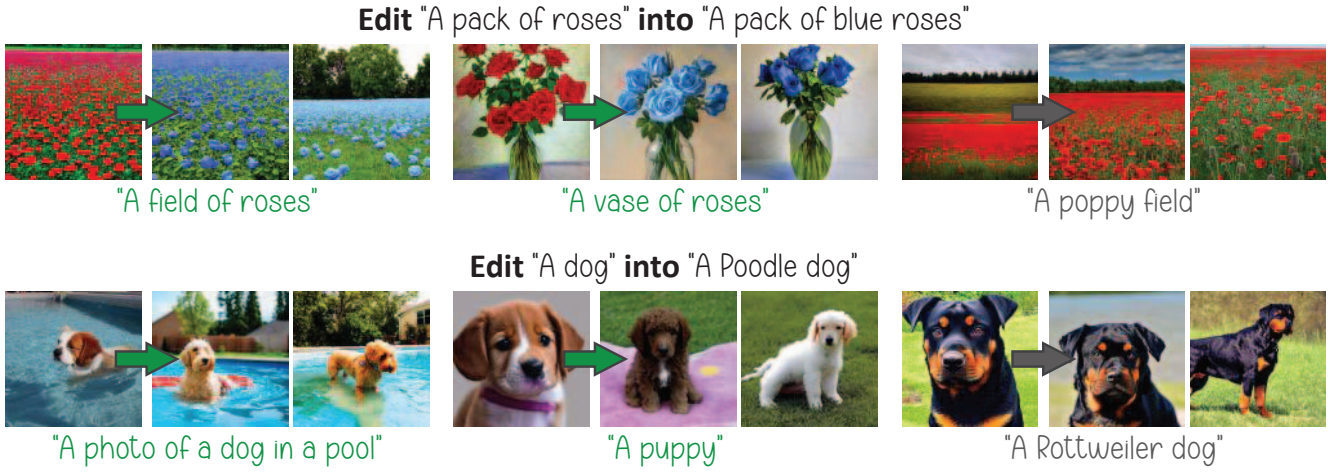

Figure 1: TIME edits implicit assumptions in a model (*e.g.*, roses are red). As a result, related prompts (green) change their behavior, while unrelated ones (gray) do not. For example, after model editing, the roses in "A field of roses" become blue.

Abstract

Text-to-image diffusion models often make implicit assumptions about the world when generating images. While some assumptions are useful (e.g., the sky is blue), they can also be outdated, incorrect, or reflective of social biases present in the training data. Thus, there is a need to control these assumptions without requiring explicit user input or costly re-training. In this work, we aim to edit a given implicit assumption in a pre-trained diffusion model. Our Text-to-Image Model Editing method, TIME for short, receives a pair of inputs: a "source" under-specified prompt for which the model makes an implicit assumption (e.g., "a pack of roses"), and a "destination" prompt that describes the same setting, but with a specified desired attribute (e.g., "a pack of blue roses"). TIME then updates the model's cross-attention layers, as these layers assign visual meaning to textual tokens. We edit the projection matrices in these layers such that the source prompt is projected close to the destination prompt. Our method is highly efficient, as it modifies a mere 2.2% *of the model's parameters in under one second. To evaluate model editing approaches, we introduce TIMED (TIME Dataset), containing* 147 *source and destination prompt pairs from various domains. Our experiments (using Stable Diffusion) show that TIME is successful in model editing, generalizes well for related prompts unseen during editing, and imposes minimal effect on unrelated generations.*[1]

1. Introduction

Text-to-image generative models have recently risen to prominence, achieving unprecedented success and popularity [54, 52, 57, 2]. The generation of high quality images based on simple textual prompts has been enabled by generative diffusion models [63, 64, 24] and large language models [51, 50]. These text-to-image models are trained on huge amounts of web-scraped image-caption pairs [61]. As a result, the models acquire implicit assumptions about the world based on correlations and biases found in the training data. This knowledge manifests during generation as visual associations to textual concepts.

Such implicit assumptions may be useful in general. For

* Equal contribution.

†Supported by the Viterbi Fellowship in the Center for Computer Engineering at the Technion.

[1] https://time-diffusion.github.io/

Figure 2: Text-to-image models make implicit assumptions on the world when generating images, as seen in the top row (*e.g.*, roses are red). In the bottom row, we override these assumptions by explicitly specifying different attributes in the prompt.

instance, the model assumes (or *knows*) that the sky is blue or that roses are red. However, in many use cases, generative model service providers may want to edit these implicit assumptions without requiring extra input from their users. Examples include updating outdated information encoded in the model (*e.g.*, a celebrity changed their hairstyle), mitigating harmful social biases learned by the model (*e.g.*, the stereotypical gender of a doctor), or generating scenarios in an alternate reality (*e.g.*, gaming) where facts are changed (*e.g.*, roses are blue). When editing such assumptions, we do not require the user to explicitly request the change, but rather aim to apply the edit directly to the model. We also generally try to avoid expensive data recollection and filtering, as well as model retraining or finetuning. These would consume considerable time and energy, thus significantly increasing the carbon footprint of deep learning research [65]. Moreover, finetuning a neural network may lead to catastrophic forgetting and a drop in performance in general [40, 34], and in model editing [78].

While text-to-image models implicitly assume certain attributes for under-specified text prompts, they can generate alternative ones when explicitly specified, as shown in Figure 2. We use this capability to replace the model's assumption with a user-specified one. Therefore, our proposed method for **T**ext-to-**I**mage **M**odel **E**diting (TIME) receives an under-specified "source" prompt, which is requested to be well-aligned with a "destination" prompt containing an attribute that the user wants to promote. While some recent work has focused on altering the model outputs for a specific prompt [19] or image [33], we target a fundamentally different objective. We aim to edit the model's *weights* such that its perception of a given concept in the world is changed. The change is expected to manifest in generated images for related prompts, while not affecting the characteristics or perceptual quality in the generation of different scenes. This would allow us to fix incorrect, biased, or outdated assumptions that text-to-image models may make.

To achieve this, we focus on the rendezvous point of the two modalities: text and image, which meet in the cross-attention layers. The importance of attention layers in diffusion models was also observed by researchers in different contexts [19, 27, 67, 7, 37]. TIME modifies the projection matrices in these layers to map the source prompt close to the destination, without substantially deviating from the original weights. Because these matrices operate on textual data irrespective of the diffusion process or the image contents, they constitute a compelling location for editing a model based on textual prompts. TIME is highly efficient: It does not require training or finetuning, it can be applied in parallel for all cross-attention layers, and it modifies only a small portion of the diffusion model weights while leaving the language model unchanged. When applied on the publicly available Stable Diffusion [54], TIME edits a mere 2.2% of the diffusion model parameters, does not modify the text encoder, and applies the edit in a fraction of a second using a single consumer-grade GPU.

For evaluating our method and future model editing efforts, we introduce a **T**ext-to-**I**mage **M**odel **E**diting **D**ataset (TIMED), containing 147 pairs of source and destination texts from various domains, as well as related prompts for each pair to assess the model editing quality. TIME exhibits impressive model editing results, generalizing for related prompts while leaving unrelated ones mostly intact. For instance, in Figure 1, requesting "a vase of roses" outputs blue roses, whereas the poppies in "a poppy field" remain red. Moreover, the generative capabilities of the model are preserved after editing, as measured by Fréchet Inception Distance (FID) [21]. The effectiveness, generality, and specificity of TIME are highlighted in subsection 5.5.

We further apply TIME for social bias mitigation, focusing on gender bias in the labor market. Consistent with concurrent work [4, 8, 15, 66], we find that text-to-image models encode stereotypes, as reflected in their image generations for professions. For instance, for the prompt "A photo of a CEO", only 4% of generated images (with random seeds) contain female figures. We edit the model to generate an image distribution that more equally represents males and females for a given profession. TIME successfully reduces gender bias in the model, improving the equal representation of genders for many professions.

To the best of our knowledge, TIME is the first method that suggests a model editing technique [12, 42] for text-to-image models. We hope that our proposed method, insights, and provided datasets will help enable future advances in text-to-image model editing, especially as these models get rapidly deployed in consumer-facing applications.

2. Related Work

Several recent and concurrent studies have considered the task of image editing using diffusion models [45, 1, 19, 44, 33, 74, 72, 75, 10]. These methods edit a given image based on a given textual prompt, each in its own technique and settings. They show impressive results in editing the properties of different objects (*e.g.*, color, style, pose) in the image by controlling different aspects of the diffusion process. A closely related application of text-to-image diffusion models is object recontextualization, where given a small number of images of an object, the goal is to generate images of the same object in different novel settings based on text prompts [16, 55, 37]. These lines of research address the tasks of editing a specific image, or generating images with novel concepts. In our work, we consider a fundamentally different objective: We aim to edit a text-to-image diffusion model's *world knowledge* using text prompts. This should cause the desired change to occur not only in the exact requested prompt, but also in generated images of related prompts. Simultaneously, unrelated generations should remain unaffected.

Editing the knowledge embedded in neural networks has been an active area of research in recent years, achieving remarkable successes in editing language models [78, 12, 11, 42, 43, 53], generative adversarial networks [3, 73, 22], and image classifiers [60]. Similar to several such techniques [3, 42, 43], our work focuses its model editing in a concise portion of the neural network.

3. Background

Denoising diffusion probabilistic models [63, 64, 24], more commonly known as diffusion models, are a family of generative models that have recently rose to prominence. They have achieved state-of-the-art performance in image generation [13, 31, 48, 29], and impressive results in downstream tasks [30, 9, 46, 17, 68, 79, 32, 59] as well as audio [36, 28, 49], video [71, 77, 23, 62], and text [18, 38] generation. Diffusion models generate their outputs using an iterative stochastic noise removal process that follows a predefined noise level schedule $\{\beta_t\}_{t=1}^{T}$. Starting from $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, in every iteration, the current sample $\mathbf{x}_t$ is denoised using a neural network $D_\theta(\mathbf{x}_t, t)$, and the next sample $\mathbf{x}_{t-1}$ is then obtained through a predefined update rule, β_t, and a stochastic noise addition. The last sample $\mathbf{x}_0$ constitutes the final synthesized output.

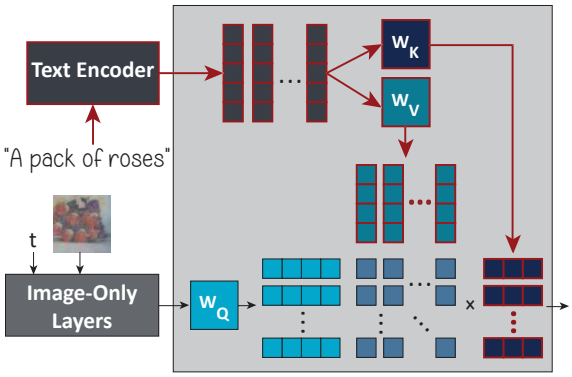

Figure 3: A cross-attention layer in a text-to-image diffusion model. We target the strictly text-based layers and the information they encode (highlighted in red).

The generative diffusion process can be controlled via additional inputs $\mathbf{c}$ to the denoising model $D_\theta(\mathbf{x}_t, t, \mathbf{c})$. The conditioning signal $\mathbf{c}$ may be a low-quality version of a desired image [58, 56], a class label [25], or a text prompt describing a desired image [54, 52, 57, 2]. In the latter case, *text-to-image diffusion models* have unveiled a new capabilaity – users can synthesize high-resolution images using simple text prompts describing the desired scenes. The remarkable success of these models has been boosted by a number of strategies, including working in a latent space [69, 54], classifier-free guidance [26], and incorporating knowledge from pre-trained text encoders such as CLIP [50] or T5 [51].

In text-to-image generation, the user-provided text prompt is input into the text encoder, which tokenizes it and outputs a sequence of token embeddings $\{\mathbf{c}_i\}_{i=1}^{l}$ describing the sentence's meaning, where $\mathbf{c}_i \in \mathbb{R}^c$. Then, in order to condition the diffusion model D_θ on them, these embeddings are injected at the cross-attention layers [14] of the model. They are projected into keys $\mathbf{K} \in \mathbb{R}^{l \times m}$ and values $\mathbf{V} \in \mathbb{R}^{l \times d}$, using learned projection matrices $\mathbf{W}_K \in \mathbb{R}^{m \times c}$ and $\mathbf{W}_V \in \mathbb{R}^{d \times c}$, respectively. The keys are then multiplied by a query $\mathbf{Q} \in \mathbb{R}^{n \times m}$, which represents visual features of the current intermediate image $\mathbf{x}_t$ in the diffusion process. This results in the following *attention map*:

$$\mathbf{M} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{m}}\right). \tag{1}$$

The attention map encodes the relevance of each textual token to each visual one. Finally, the cross-attention output is calculated as

$$\mathbf{O} = \mathbf{M}\mathbf{V}, \tag{2}$$

which constitutes a weighted average of all textual values for each visual query. This output then propagates to the subsequent layers of the diffusion model D_θ. The cross-attention mechanism is visually depicted in Figure 3. Its

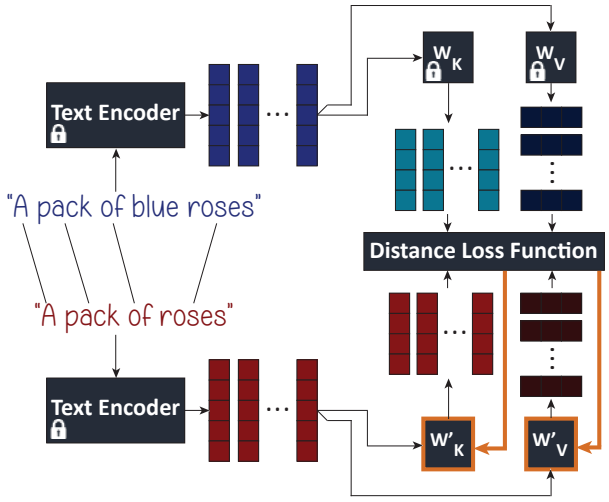

Figure 4: An overview of TIME. $\mathbf{W'}_K$ and $\mathbf{W'}_V$ are edited to map the source prompt's embeddings close to the destination prompt's keys and values. The loss is regularized for specificity.

expressiveness is increased by using multi-headed attention [70], and by incorporating it in multiple layers in the model architecture.

4. TIME: Text-to-Image Model Editing

We propose an algorithm for **T**ext-to-**I**mage **M**odel **E**diting (TIME). Our algorithm takes two textual prompts as input: an under-specified *source prompt* (*e.g.*, "a pack of roses"), and a similar more specific *destination prompt* (*e.g.*, "a pack of **blue** roses"). We aim to shift the source prompt's visual association to resemble the destination.

To this end, we focus on the layers that map textual data into visual data – the cross-attention layers. In each such layer, the matrices $\mathbf{W}_K$ and $\mathbf{W}_V$ project the text embeddings into keys and values that the visual data attends to. Because these keys and values are computed independently of the current diffusion step or image data, we identify them as the knowledge editing targets (see Figure 3).

Let $\{\mathbf{c}_i\}_{i=1}^{l}$ and $\{\mathbf{c'}_j\}_{j=1}^{l'}$ be the source and destination prompt's embeddings, respectively. For each source embedding $\mathbf{c}_i$ stemming from a token $\mathbf{w}_i$ (*e.g.*, "roses" in "a pack of roses"), we identify the destination embedding that corresponds to the same token, and denote it as $\mathbf{c}_i^*$. Note that embeddings stemming from additional tokens in the destination prompt (*e.g.*, "blue" in "a pack of blue roses") are discarded. Nevertheless, their influence is present in other destination tokens through the text encoder architecture.

In each cross-attention layer in the diffusion model, we calculate the keys and values of the destination prompt as

$$\begin{aligned} \mathbf{k}_i^* &= \mathbf{W}_K \mathbf{c}_i^*, \qquad \text{for } i = 1, \ldots, l, \\ \mathbf{v}_i^* &= \mathbf{W}_V \mathbf{c}_i^*, \qquad \text{for } i = 1, \ldots, l. \end{aligned} \tag{3}$$

We then optimize for new projection matrices $\mathbf{W'}_K$ and $\mathbf{W'}_V$ that minimize the following loss function:

$$\begin{aligned} &\sum_{i=1}^{l} \|\mathbf{W'}_K \mathbf{c}_i - \mathbf{k}_i^*\|_2^2 + \lambda \|\mathbf{W'}_K - \mathbf{W}_K\|_F^2 \\ &+ \sum_{i=1}^{l} \|\mathbf{W'}_V \mathbf{c}_i - \mathbf{v}_i^*\|_2^2 + \lambda \|\mathbf{W'}_V - \mathbf{W}_V\|_F^2, \end{aligned} \tag{4}$$

where $\lambda \in \mathbb{R}^+$ is a hyperparameter, $\|\cdot\|_2$ is the ℓ_2 norm, and $\|\cdot\|_F$ is the Frobenius norm. This loss function encourages the source prompt generation to behave similarly to the destination prompt generation, while preserving proximity to the original projection matrices. Note that this loss function (depicted in Figure 4) can be minimized for each cross-attention layer in a completely parallel and independent manner. Moreover, as we prove in the supplementary material, the loss has a closed-form global minimum at

$$\begin{aligned} \mathbf{W'}_K &= \left(\lambda \mathbf{W}_K + \sum_{i=1}^{l} \mathbf{k}_i^* \mathbf{c}_i^\top\right) \left(\lambda \mathbf{I} + \sum_{i=1}^{l} \mathbf{c}_i \mathbf{c}_i^\top\right)^{-1}, \\ \mathbf{W'}_V &= \left(\lambda \mathbf{W}_V + \sum_{i=1}^{l} \mathbf{v}_i^* \mathbf{c}_i^\top\right) \left(\lambda \mathbf{I} + \sum_{i=1}^{l} \mathbf{c}_i \mathbf{c}_i^\top\right)^{-1}. \end{aligned} \tag{5}$$

Finally, we use the modified text-to-image diffusion model with the new projection matrices to generate images. We expect this modified model to comply with the new assumption requested by the user.

We experiment with different versions of the loss function in Equation 4 (*e.g.*, only editing $\mathbf{W'}_V$, varying λ) and show this ablation study in the supplementary material.

5. Experiments

5.1. Implementation Details

We use the publicly available Stable Diffusion [54] version 1.4 as the backbone text-to-image model, with its default hyperparameters. This model contains 16 cross-attention layers, whose key and value projection matrices constitute a mere 2.2% of the diffusion model parameters. TIME edits these matrices in around 0.4 seconds using a single NVIDIA RTX 3080 GPU. We use $\lambda = 0.1$ and utilize augmented versions of the source and destination text prompts while editing, in line with the findings of the ablation study in the supplementary material.

We also provide the full set of hyperparameters and our code in the supplementary material. Note that λ is chosen differently when mitigating social biases, as explained in section 6.

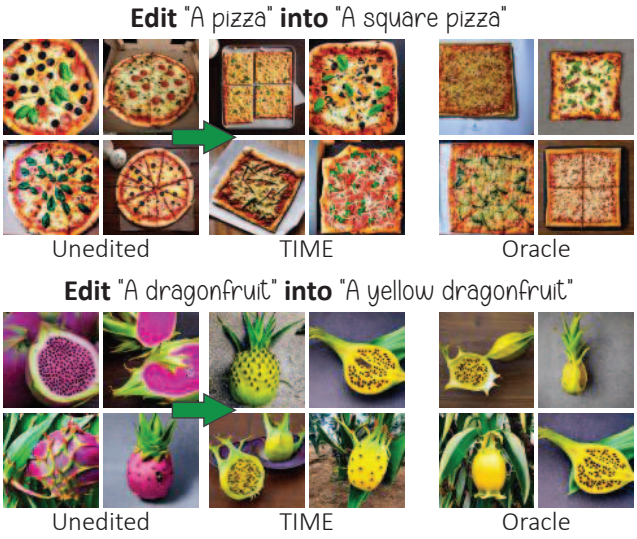

Figure 5: Using TIME, image generations for the source prompt mimic the the destination prompt's oracle behavior.

Editing	Source	Destination
	A dog	A green dog
Testing	**Source**	**Destination**
Positives	A puppy	A green puppy
	An angry dog	A green angry dog
	A bulldog	A green bulldog
	A chihuahua	A green chihuahua
	A pixel art of a dog	A pixel art of a green dog
Negatives	A cat	A green cat
	A bunny	A green bunny
	A hyena	A green hyena
	A fox	A green fox
	A wolf	A green wolf

Table 1: An example of a single dataset entry in TIMED.

5.2. TIME Dataset

To establish an evaluation benchmark for our task, we curate a **T**ext-to-**I**mage **M**odel **E**diting **D**ataset (TIMED) containing 147 entries. See Table 1 for a sample entry. Each entry in the dataset contains a pair of source and destination prompts, which are used for model editing. The source prompt (*e.g.,* "A dog") is an under-specified text prompt that describes a certain scenario in which some visual attribute is implicitly inferred by the text-to-image model. The destination prompt (*e.g.,* "A green dog") describes the same scene, but with a desired specified attribute. Additionally, each entry contains five positive prompts, for which we expect our edit to generalize (*e.g.,* "A puppy" should generate a green puppy), and five negative prompts which are semantically adjacent, but should not be affected by the edit (*e.g.,* "A cat" should not generate a green cat). Each positive or negative prompt is associated with its own destination prompt for evaluation purposes. Positive prompts are expected to gravitate towards their destination prompt, whereas negative ones should not. The dataset contains a wide variety of implicit assumptions to edit from different domains. We additionally compile a smaller disjoint validation set, which we use for hyperparameter tuning.

To ensure a valid evaluation on Stable Diffusion [54] v1.4, we filter out test set entries for which the unedited model shows poor generative quality, retaining 104 examples. The full dataset and filtering process are provided in the supplementary material.

Figure 6: TIME generalizes to prompts related to the input (left), with minimal effect on unrelated ones (right).

5.3. Qualitative Evaluation

As we show in Figure 5, TIME successfully edits the behavior of the diffusion model for the provided source prompt. Moreover, our method can generalize for related

Figure 7: Generation results on a positive (green) and negative (gray) prompt for the same edit under different λ values. As λ increases, we trade off generality (paintings of cows being on a beach) for specificity (goats being on a beach).

text prompts with minimal effect on unrelated ones, as highlighted in Figures 1, 6, and the supplementary material.

When editing a model based on a given text prompt, we need to control the extent to which the edit affects other prompts. Therefore, there exists a natural trade-off between generality and specificity, as we demonstrate in Figure 7.

5.4. Evaluation Metrics

To accurately assess the performance of our text-to-image model editing technique, we focus on three concepts set forth by efforts in language model editing literature [43]: efficacy, generality and specificity. **Efficacy** measures how effective the editing method is on the source prompt used for editing. **Generality** measured how the editing method generalizes to other related prompts, using the positive test prompts in TIMED. **Specificity** measures the ability to leave the generation of unrelated prompts unaffected, using the negative test prompts in TIMED.

For each source test prompt in each TIMED entry, we generate 24 images using different random seeds. We use CLIP [50] to classify images generated with the source prompt as either the source or destination text, and then compute the fraction of images classified as the desired option – the destination prompt for efficacy and generality, and the source prompt for specificity. We report average metrics along with standard deviations across random seeds.

Furthermore, to evaluate the effect of TIME on the overall generative quality of the model, we report Fréchet Inception Distance (FID) [21] and CLIP Score [20] on MS-COCO [39], following standard practice [54, 57, 52, 2]. See supplementary material for more details on the metrics.

	Oracle	Baseline	TIME
Efficacy (↑)	98.08% ±01.10	10.50% ±03.27	88.10% ±02.85
Generality (↑)	94.72% ±01.21	12.33% ±01.11	69.04% ±02.15
Specificity (↑)	90.13% ±01.50	90.13% ±01.50	68.34% ±02.07
FID (↓)	12.67	12.67	12.10
CLIP Score (↑)	31.24	31.24	30.88

Table 2: Evaluation results on 104 TIMED test set entries. Efficacy, generality, and specificity assess the model editing quality. FID and CLIP Score measure the generative quality on the MS-COCO dataset [39].

5.5. Quantitative Evaluation

We report the results of a *baseline*, which refers to the unedited model's results using the source prompt for all generations. We also define an *oracle*, which is the same unedited model using the destination positive prompts (which are unavailable to TIME) for the positive samples and the source negative prompts for the negative samples. The oracle serves as an upper bound for the potential performance of model editing techniques based on text inputs.We also experimented with model finetuning. Results are shown in the supplementary material.

We summarize our results in Table 2. As the first text-to-image model editing technique, TIME shows promising

Figure 8: TIME debiases a text-to-image model, making it equally represent genders in test prompts for "A developer".

results. In addition to its high efficacy, TIME is able to generalize to many related prompts. As expected, the edited model generates the desired concept substantially more often than the baseline model. While the model sustains a drop in specificity, its overall generative quality remains unaffected. This is verified by the FID [21] and CLIP Score [20] metrics on the MS-COCO [39] dataset, which are comparable to the baseline unedited model.

While we use a fixed λ for Table 2, different editing scenarios would benefit from tuning λ in accordance with their needs. See supplementary material for further discussion and experiments with the generality–specificity trade-off.

In this work, we concentrate on editing a single assumption at a time. For preliminary experiments with editing multiple assumptions, see supplementary material.

6. TIME for Gender Bias Mitigation

In the previous section, we evaluated TIME for editing implicit model assumptions. In this section, we address social bias as a particular case of implicit assumptions made by the model. It is well-documented that language models [5, 6, 76] and text-to-image diffusion models [4, 8, 15, 66] implicitly encode social and cultural biases.

For instance, models assume a certain stereotypical gender based on a person's profession (*e.g.*, only 4.0% of images generated for "A photo of a CEO" contain female figures). This may lead to the perpetuation of existing stereotypes [41], as these models are rapidly deployed in a variety of applications (*e.g.*, marketing, media). Therefore, we aim to *erase* the assumptions that encode stereotypes, rather than *edit* them, such that the model will not make any (possibly harmful) assumptions.

While many types of social biases exist, we consider gender bias within the labor market as a case study. To this end, we address the male–female inequality in the portrayal of different professions. We acknowledge that our current perspective is narrow since it only considers binary genders and may exclude and marginalize non-binary individuals. However, we also recognize the risk of introducing other, unwanted stereotypes regarding the visual features of non-binary genders. We look forward to future research that can better incorporate more gender identities with detailed and carefully defined data.

6.1. Data Preparation

We compose a dataset of 35 entries with under-specified source prompts of the form "A/An [profession]", such as "A CEO". We identify the stereotypical gender for each such profession using a list compiled by [76], based on United States labor force statistics. The destination prompt is then defined as "A [gender] [profession]" using the non-stereotypical gender, such as "A female CEO". In order to evaluate our debiasing efforts, we further include five test prompts for each profession describing it in different scenarios, *e.g.*, "A CEO laughing". We make the dataset publicly available and and provide more details about it in the supplementary material.

6.2. Method Description

For each profession p, we aim for 50% of generations to be female and 50% to be male. We control the strength of the debiasing by tuning λ (from Equation 4). Smaller λ values steer the model towards the non-stereotypical gender, whereas larger ones encourage it to maintain its existing assumptions. Note that as the baseline model is more biased, the editing should be stronger. Consequently, we binary search for a different λ_p for each profession p, aiming for an equal gender representation in generations for the

		Baseline	Oracle	TIME	TIME (Multi)
Δ ($\downarrow$)		0.57 $\pm$ 0.011	0.142 $\pm$ 0.084	0.28 $\pm$ 0.002	0.48 $\pm$ 0.015
F_p	Hairdresser	72.00%	50.00%	53.60%	66.67%
	CEO	04.00%	50.00%	35.20%	33.33%
	Teacher	84.80%	50.00%	35.20%	25.00%
	Lawyer	28.80%	55.83%	61.60%	50.00%
	Housekeeper	99.20%	47.50%	56.00%	83.33%
	Farmer	02.40%	48.33%	49.59%	33.33%

Table 3: Gender bias results for the baseline model, and after debiasing using TIME. The metrics are calculated over the test prompts, which are unseen during editing.

validation prompt "A photo of a/an [profession]".

6.3. Gender Bias Estimation

To measure the degree of gender inequality in a text-to-image model's perception of a profession, we estimate the percentage of female figures generated by it for each profession, denoted as $F_p \in [0, 100]$. To do so, we generate 24 images for each test prompt, and use CLIP [50] to classify gender in each image. We then determine the normalized absolute difference between the observed percentage F_p and the desired gender equality for a profession p, represented by $\Delta_p = |F_p - 50|/50$. To obtain a single comprehensive measure of gender bias within the model, we compute the average value of Δ_p across all professions in the dataset and denote it as Δ. An ideal, unbiased model should satisfy $\Delta = 0$.

6.4. Results

Our results are summarized in Table 3. We present Δ, along with the percentage of females F_p in the test prompt generations for a representative subset of professions. We report these metrics for various models. The *baseline* model stands for the unedited model's bias. The *oracle* is defined as the unedited model when prompted with an explicit prompt of the form "a [gender] [profession]", where [gender] is randomized in each generation to be either "famale" or "male". We also perform a multi-assumption editing experiment, **TIME (Multi)**, where a single λ is chosen based on the validation set for debiasing all professions at once.

TIME successfully reduces the bias metric Δ to less than a half of the baseline model's bias. When carefully examining our results, some professions, such as hairdresser and CEO, become less biased by attaining a more equal gender distribution. Others, such as teacher and lawyer, become anti-biased (*i.e.*, biased towards the non-stereotypical gender). Moreover, some professions, such as housekeeper and farmer, are effectively debiased to almost equally represent both females and males. After using TIME, 14 professions exhibit a low test prompt bias metric $\Delta_p \in [0, 0.2]$, representing near-optimal equality. In contrast, only 8 professions displayed such behavior in the baseline model. Moreover, the choice of prompt affects the observed ratio, as discussed in the supplementary material. We also note that although the oracle serves as an upper bound for debiasing, using the oracle as a debiasing method in a production system may not easily generalize and require further adjustments. However, debiasing with TIME is able to generalize and adapt to different prompts – see Figure 9.

While TIME with multi-editing is also successful at reducing bias, it is less effective. Debiasing multiple professions at once is difficult because debiasing one profession affects on the gender ratio of other professions, as can be observed in Figure 10. Interestingly, professions that share the same stereotypical gender tend to have a stronger effect on one another. For example, when we edit software-developer prompts to generate more female figures, we also cause CEO prompts to generate more female figures. While it is debatable whether this effect is desired or not, it causes the debiasing of multiple professions to be trickier to control. We leave this issue to be investigated in future work, perhaps by expanding TIME. Moreover, further investigating specificity, we found that editing "a/an [profession]" towards male direction does not hurt the generation of "a female [profession]", as it produces 100% female figures pre-edit and 99.7% post-edit, with similar results for editing towards female (94% vs. 88.4%).

7. Limitations

While recent advances in text-to-image generative modelling have shown great performance, these models may fail to generate images aligned with the requested prompts in some cases, such as compositionality or counting [52, 57, 47]. TIME aims to edit the assumptions in the model for a user-specified prompt. It is not designed to teach the model new visual concepts that it was unable to generate. Thus, TIME inherits the generative limitations of the model

Figure 9: After debiasing "physician" with TIME, it generalizes to related professions (neurologist, surgeon) while adapting to gendered prompts: it produces only female figures for "a pregnant doctor". An oracle baseline will not be able to perform the same.

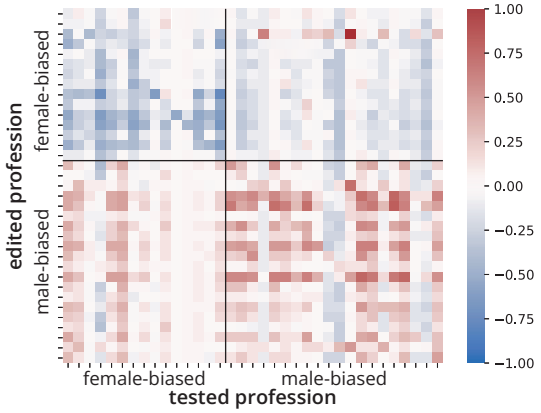

Figure 10: Effect of debiasing one profession on other professions. Values denote F_p in the generated image.

it edits, as evident in the Pearson correlation coefficient between the oracle generative performance and TIME's success, $\rho = 0.73$. This strongly suggests that TIME is more likely to succeed when the oracle model successfully generates the desired concepts.

Moreover, as shown in Figure 11, TIME sometimes applies an edit too mildly (hindering generality) or too aggressively (hindering specificity). Future work may address this limitation by devising algorithms for automatically adjusting λ on a per-edit basis, or via alternative regularization methods that improve the generality–specificity tradeoff.

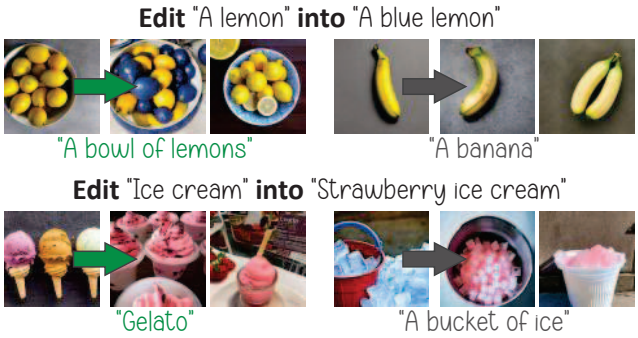

Figure 11: In some cases, TIME applies the requested edit too mildly (top), or too aggressively (bottom).

8. Conclusion

In this work, we propose the following research question: How can specific implicit assumptions in a text-to-image model be edited after training? To investigate this question, we present TIME, a method that explores this task. TIME edits models efficiently, and produces impressive results. We additionally introduce a dataset, TIMED, for evaluating text-to-image model editing methods. As text-to-image generative models get deployed in consumer-facing applications, methods for quickly editing the associations and biases embedded in them are important. We hope that our method and datasets will help pave the way for future advances in text-to-image model editing.

This work can be expanded in many possible directions. One direction is to analyze the role of different components in storing and retrieving knowledge: different elements of the cross-attention mechanism and different tokens in the prompt. It would also be interesting to expand the method for editing multiple facts in bulk while maintaining the model's performance. We presented evidence that TIME is able to reduce gender bias, and it would be beneficial to further investigate this direction towards a more comprehensive debiasing method.

Acknowledgements

This research was supported by the Israel Science Foundation (grant No. 448/20), an Azrieli Foundation Early Career Faculty Fellowship, an AI Alignment grant from Open Philanthropy, a grant from the FTX Future Fund regranting program, the Crown Family Foundation Doctoral Fellowship, and the Israeli Council For Higher Education – Planning & Budgeting Committee. We also thank Dana Arad, Roy Ganz, Itay Itzhak, and Gregory Vaksman for their valuable feedback and discussions on this work.