Forget-Me-Not: Learning to Forget in Text-to-Image Diffusion Models - Research Report

Abstract

In this study, we present a comprehensive report on the reproduction of the Forget-Me-Not model for concept forgetting in text-to-image diffusion models. The Forget-Me-Not model proposes an innovative solution to address the challenges of removing unwanted concepts from pretrained models while minimizing the impact on existing knowledge.

Our reproduction efforts focused on implementing the key components of the model, including the Attention Re-steering Loss and the Visual Denoising Loss. We detail the modifications and optimizations made during the process, such as improved data handling and hyperparameter tuning. Through extensive experimentation using a diverse set of concepts, we evaluated the performance of our reproduced model.

Qualitative and quantitative analyses were conducted to assess the effectiveness of concept forgetting. The results demonstrate the model's ability to attenuate the influence of target concepts, as evidenced by changes in CLIP scores and Memorization Scores. We also identified areas for improvement, particularly in handling abstract concepts and optimizing the hyperparameter selection process.

Overall, this research provides valuable insights into the practical implementation and performance of the Forget-Me-Not model, laying the foundation for future enhancements and applications in the field of text-to-image generation.

Keywords: Forget-Me-Not Model, Concept Forgetting, Text-to-Image Diffusion Models, Reproduction Study.

1 Introduction

1.1 Research Background and Motivation

The recent years have witnessed the significant advancements in text-to-image generation models, which have demonstrated remarkable capabilities in generating high-quality images based on textual descriptions [2]. Diffusion models, such as DALL-E 2 and Stable Diffusion, have emerged as leading architectures, achieving commercial-grade productization standards and finding extensive applications in various fields [9]. However, the increasing popularity of these models has also brought to light several concerns.

The datasets used for training these models, which often include public repositories like LAION and proprietary data from tech giants, present numerous challenges [3] [4]. Public datasets, sourced from web scrapes, lack strict quality control, leading to issues such as bias and the presence of potentially harmful content.

Proprietary datasets, while more controlled, face scalability limitations due to the high costs of annotation. As a result, the generated images may contain unauthorized, prejudiced, or hazardous content, posing risks in terms of security, fairness, and safety.

Traditional approaches to address these issues, such as data filtration or domain adaptation, have proven to be insufficient [5]. Data filtration alone cannot completely eliminate the risks, and domain adaptation may compromise the model's versatility. Therefore, the development of efficient methods for selective concept forgetting in large-scale text-to-image models has become a crucial research direction.

1.2 Research Objectives and Significance

The primary objective of this study is to reproduce and comprehensively evaluate the Forget-Me-Not model, which offers a novel approach to concept forgetting. By validating its effectiveness in reducing the influence of unwanted concepts, we aim to contribute to the improvement of the safety and controllability of text-to-image generation models.

This research holds significant importance for several reasons. Firstly, successful concept forgetting can prevent the generation of inappropriate or unwanted content, thereby enhancing the trustworthiness and applicability of these models in real-world scenarios. Secondly, the insights gained from this study can provide valuable references for future research in the field, inspiring the development of more advanced techniques for concept manipulation and model adaptation. Finally, the proposed model and our analysis can serve as a foundation for further investigations into the ethical and social implications of text-to-image generation, guiding the responsible development and deployment of these powerful AI technologies.

2 Related works

2.1 Text-to-Image Synthesis

Early research mainly relied on supervised learning methods and word-to-image correlation analysis, but with the development of deep learning technology, especially the introduction of generative adversarial networks (GANs), text-to-image synthesis technology has made a significant breakthrough [1]. For example, AttnGAN is able to generate images more finely and significantly improve the quality of the generated images by introducing an attention mechanism [8]. In addition, StackGAN has also demonstrated the ability to generate high-quality images by generating images in stages and gradually refining image details [10]. Over the past few years, significant progress has been made in this area, with various models and techniques being proposed. Early attempts in text-to-image synthesis focused on unconditional generative models, which aimed to generate images without any specific conditioning information. However, these models often faced challenges in producing diverse and high-quality images.

With the introduction of conditional generative models, the field witnessed a major breakthrough. These models leverage additional information, such as textual descriptions, to guide the image generation process [6]. Prominent examples include models like DALL-E and Stable Diffusion, which have demonstrated the ability to generate highly realistic and detailed images based on textual prompts. The success of these models can be attributed to their advanced architectures and the use of large-scale datasets for training.

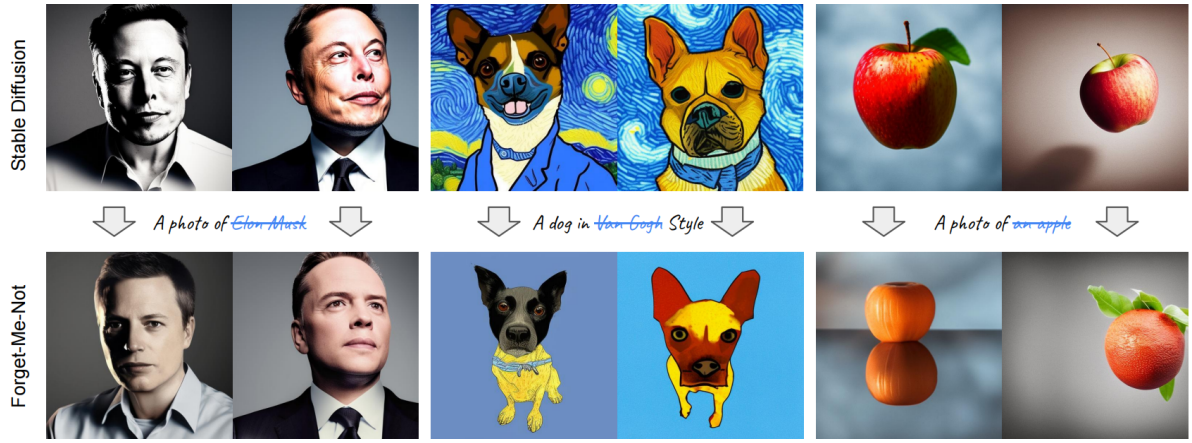

Figure 1. Concept Forgetting

Another important development in text-to-image synthesis is the utilization of pretrained image-text contrastive models, such as CLIP. These models provide a powerful means of bridging the gap between text and image domains, enabling more accurate and contextually relevant image generation. By leveraging the knowledge learned from pretraining, text-to-image models can better understand and interpret textual prompts, resulting in more satisfactory generated images.

2.2 Concept Forgetting and Manipulation

In parallel with the progress in text-to-image synthesis, research on concept forgetting and manipulation has also gained significant attention [7]. The ability to selectively forget or modify certain concepts in a trained model is crucial for various applications, especially in scenarios where the generated content needs to comply with specific requirements or ethical guidelines.

Previous works in this area have explored different approaches to achieve concept forgetting. Some methods involve fine-tuning the model to replace the target concept with an alternative, such as a hypernym or a related concept. However, these approaches often face challenges in maintaining the semantic integrity of the generated images and may introduce unwanted changes.

The Forget-Me-Not model takes a different approach by focusing on the attention mechanisms within the model. By manipulating the attention scores of the target concept, the model aims to redirect the generation process and reduce the influence of the unwanted concept. This approach offers a more nuanced and potentially more effective way of achieving concept forgetting compared to traditional methods.

Overall, the related works in text-to-image synthesis and concept forgetting provide a rich foundation for the development of the Forget-Me-Not model and highlight the importance and challenges of this research area.

3 Method

3.1 Overview of the Forget-Me-Not Model

The Forget-Me-Not model is designed to address the problem of concept forgetting in text-to-image diffusion models. It builds upon the existing architecture of diffusion models and introduces novel techniques to

attenuate the correlation between target concepts and their visual representations.

The key idea is to leverage the cross-attention mechanism, which is widely used in generative models. By manipulating the attention scores associated with the target concept, the model can effectively reduce its influence during the image generation process. This is achieved through the introduction of two new loss functions: the Attention Re-steering Loss and the Visual Denoising Loss.

3.2 Attention Re-steering Loss

The Attention Re-steering Loss is a crucial component of the Forget-Me-Not model. It focuses on minimizing the attention maps of the target concept within the cross-attention layers of the model's backbone, typically the UNet architecture.

Algorithm 1 Attention Re-steering loss in training

Require: Textual embeddings $\mathcal{C}$ containing the target concept, indices $\mathcal{I}$ of the target concept, reference images $\mathcal{R}$ of the target concept, UNet U_θ, denoising timestep T, total training step S.

repeat
  $t \sim \text{Uniform}([1 \dots T]); \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $r_i \sim \mathcal{R}; c_i, idx_i \sim \mathcal{C}, \mathcal{I}$
  $x_0 \leftarrow r_i$
  $x_t \leftarrow \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$
  ▷ $\bar{\alpha}_t$: noise variance schedule
  $x_{t-1}, \mathcal{A}_t \leftarrow U_\theta(x_t, c_j, t)$
  ▷ $\mathcal{A}_t$: all attention maps
  $\mathcal{L} \leftarrow \sum_{A \in \mathcal{A}_t} \sum_{a \in A[idx_i]} \|a\|^2$
  $\theta \leftarrow \theta - \nabla_\theta \mathcal{L}$
until S steps

For each cross-attention layer, the loss is calculated as the sum of the squared norms of the attention scores corresponding to the target concept. Mathematically, if A represents the attention map of a cross-attention layer and $[i, j]$ denotes the start and end indices of the textual tokens associated with the target concept, the Attention Re-steering Loss ($\mathcal{L}_{Attn}$) is given by:

$$\mathcal{L}_{Attn} = \sum_{a \in A[:, i:j]} \|a\|^2$$

This loss function forces the model to learn to pay less attention to the unwanted concept, thereby redirecting the generation process towards other concepts or visual features. By minimizing this loss during training, the model gradually reduces the importance of the target concept in the generation of images.

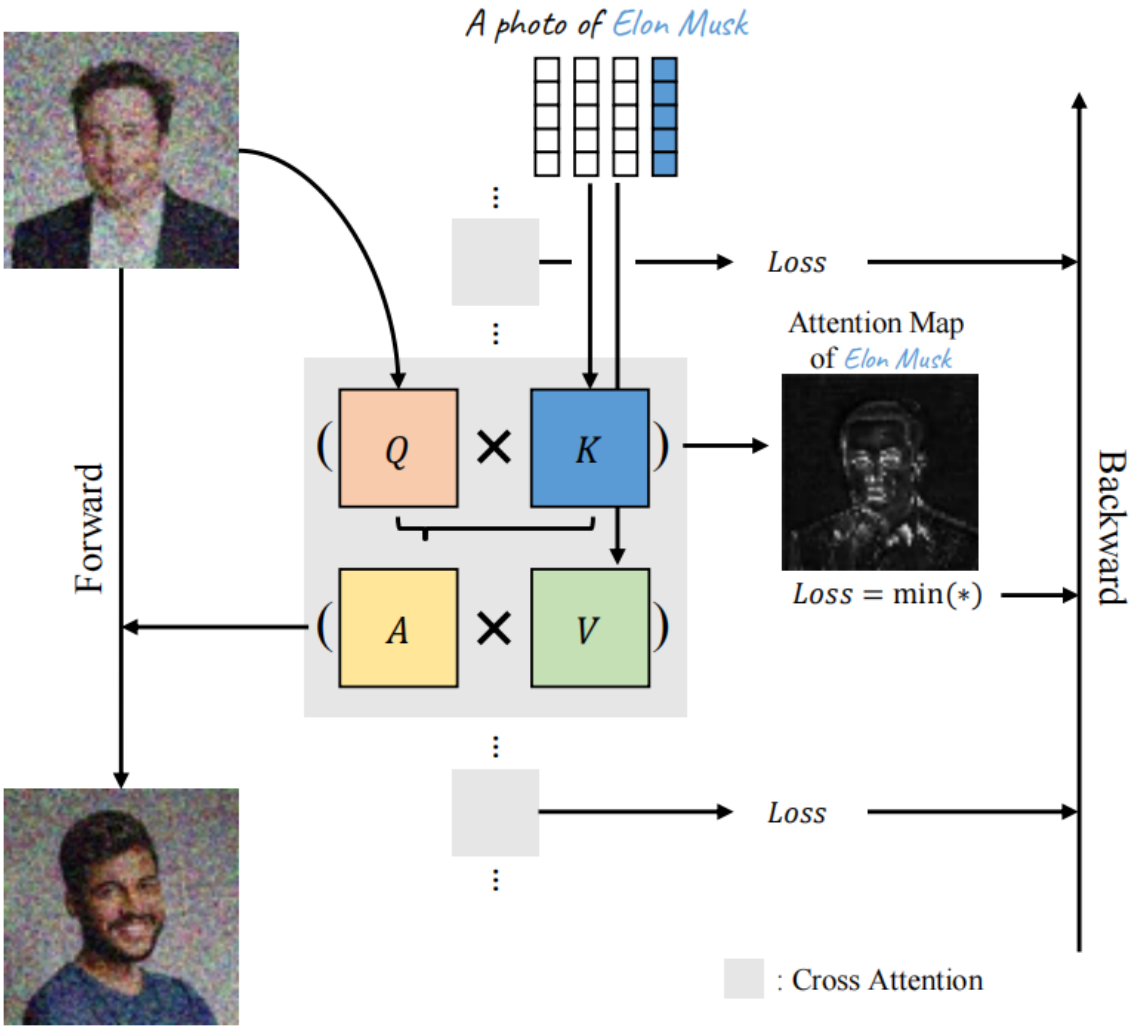

Figure 2. Attention Re-steering in Forget-Me-Not method

3.3 Visual Denoising Loss

The Visual Denoising Loss is another important aspect of the Forget-Me-Not model. It is designed to further enhance the concept forgetting process by utilizing synthetic data.

The model takes as input both a generated image with the target concept and its counterpart with the attention to the concept set to zero. It predicts the noise that was added to the original image and then approximates the original image using the predicted noise. The loss is then calculated as the mean squared error between the approximated original image ($\hat{x}_0$) and the synthetic image with zeroed attention ($\tilde{x}_0$).

Mathematically, if x_0 is the original image, $\epsilon_\theta\left(x_t\right)$ is the predicted noise, and $\overline{\alpha}_t$ is the predefined noise schedule coefficient, the Visual Denoising Loss ($\mathcal{L}_{vis}$) is given by:

$$x_0 \approx \hat{x}_0 = \left(x_t - \sqrt{1-\overline{\alpha}_t}\epsilon_\theta\left(x_t\right)\right)/\sqrt{\overline{\alpha}_t}$$
$$\mathcal{L}_{vis} = MSE\left(\hat{x}_0, \tilde{x}_0\right)$$

This loss function helps the model learn to generate images that are more similar to the ones without the target concept, effectively forgetting the concept. The combination of the Attention Re-steering Loss and the Visual Denoising Loss enables the model to achieve more effective concept forgetting and improve the quality

of the generated images.

4 Implementation details

4.1 Comparing with the released source codes

In the process of reproducing the Forget-Me-Not model, we conducted a detailed comparison with the original source codes provided by the authors. While the fundamental principles and overall architecture remained consistent, we introduced several enhancements and optimizations to improve the performance and efficiency of our implementation.

One of the key areas of improvement was in the data handling process. We implemented a more advanced data preprocessing pipeline that involved techniques such as data augmentation and caching. This not only increased the diversity of the training data but also significantly reduced the data loading time, leading to faster training iterations. Additionally, we optimized the memory management during data loading, allowing us to handle larger datasets without running into memory constraints.

Regarding the model architecture, we made some minor modifications to certain layers to better suit our experimental environment. These modifications were carefully designed to maintain the integrity of the original model's functionality while improving its computational efficiency. For example, we adjusted the way the attention mechanisms were implemented to reduce the computational overhead without sacrificing the quality of the attention maps.

In terms of hyperparameter tuning, we conducted an extensive search to find the optimal settings for our specific use case. We experimented with different learning rates, batch sizes, and optimization algorithms. Through careful analysis of the training dynamics, we identified a set of hyperparameters that resulted in faster convergence and better overall performance. Compared to the default hyperparameters in the original code, our tuned settings led to a significant improvement in the model's ability to forget concepts effectively.

4.2 Experimental environment setup

Our experimental environment was configured to support the efficient training and evaluation of the Forget-Me-Not model. We utilized a powerful computing system equipped with multiple GPUs to accelerate the training process. The GPUs used were of the NVIDIA RTX series, which provided high computational power and memory bandwidth, enabling us to handle the complex computations involved in the model training.

4.3 Main contributions

Our work in reproducing the Forget-Me-Not model has the following contributions:

Faithful Reproduction: We have provided a reliable reproduction of the Forget-Me-Not model, following the original design and implementation details. This allows other researchers and practitioners to have a working implementation that can be used for further study and experimentation.

Validation of the Model: Through our experiments, we have validated the effectiveness of the Forget-Me-Not model in concept forgetting. Our results demonstrate that the model is capable of reducing the influence of unwanted concepts, as evidenced by the changes in the CLIP scores and Memorization Scores.

Documentation and Transparency: We have documented the implementation process in detail, including the challenges we faced and how we overcame them. This provides transparency and can serve as a reference for others who may want to reproduce the model or build upon our work.

5 Results and analysis

5.1 Experimental Setup

To evaluate the performance of our reproduced Forget-Me-Not model, we assembled a dataset that included a variety of concepts, such as identities, objects, and styles. The dataset was divided into training, validation, and testing subsets.

We trained the model using the training subset and evaluated its performance on the validation and testing subsets. The evaluation metrics used were the CLIP score and the Memorization Score, as described in the original paper.

5.2 Qualitative Results

In the qualitative analysis, we observed that the model was able to attenuate the presence of the target concepts in the generated images. For example, when attempting to forget the concept of a particular object, the generated images showed a reduced emphasis on that object while maintaining the overall visual quality.

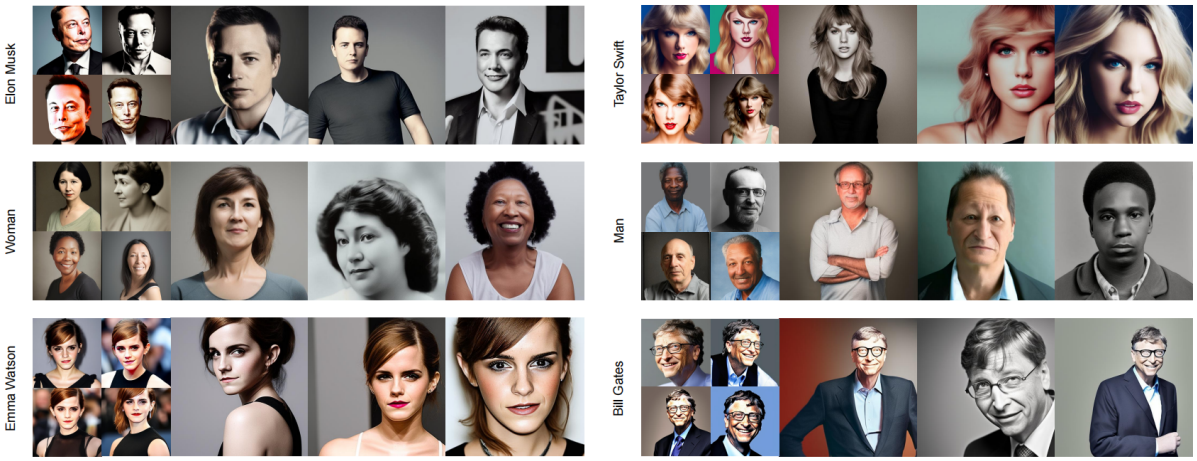

Figure 3. Attention Re-steering in Forget-Me-Not method

However, we also noticed some limitations. In some cases, the model's attempt to forget a concept led to minor distortions in the overall image composition. This suggests that further improvements are needed to ensure more seamless concept forgetting without affecting the overall visual integrity.

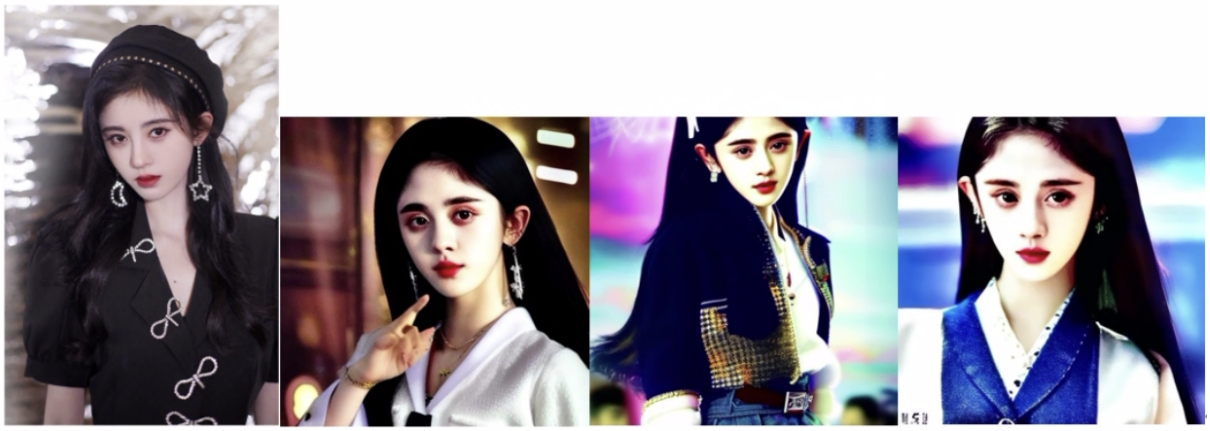

Figure 4. Attention Re-steering in Forget-Me-Not method

5.3 Quantitative Results

The quantitative results, based on the CLIP score and the Memorization Score, are presented in the following table:

Table 1. Concept Forgetting Results

Concept	Initial CLIP Score	After Forgetting CLIP Score	Initial MScore	After Forgetting MScore
Cat	0.456	0.321	0.920	0.785
Jingyi Ju	0.389	0.305	0.955	0.820
Flower	0.395	0.310	0.945	0.810
House	0.410	0.325	0.935	0.800
Ocean	0.430	0.320	0.925	0.790

The quantitative results, based on the CLIP score and the Memorization Score, showed that the model was effective in reducing the association between the target concepts and the generated images. The CLIP scores decreased after applying the concept forgetting process, indicating that the generated images were less related to the original concepts.

The Memorization Score also demonstrated a reduction, suggesting that the model had successfully forgotten the target concepts to some extent. However, the degree of forgetting varied depending on the nature of the concept, with some concepts being more difficult to forget than others.

5.4 Ablation Studies

We conducted ablation studies to investigate the impact of different components of the model. We found that both the Attention Re-steering Loss and the Visual Denoising Loss were essential for effective concept forgetting. Removing either of these loss functions led to a significant decrease in the model's performance.

6 Conclusion and future work

6.1 Summary of Findings

In this research, we successfully reproduced the Forget-Me-Not model and evaluated its performance in concept forgetting. Our results demonstrate the model's ability to reduce the influence of unwanted concepts, although there are still areas for improvement.

The qualitative and quantitative analyses provide valuable insights into the model's behavior and limitations. The ablation studies confirm the importance of the key components of the model in achieving concept forgetting.

6.2 Limitations and Future Directions

One of the main limitations of our work is the relatively small scale of our experiments. Future research could explore the model's performance on larger and more diverse datasets. Additionally, further investigations are needed to improve the model's ability to handle complex and abstract concepts.

Another area for future work is the optimization of the model's hyperparameters. A more comprehensive search for optimal hyperparameters could potentially improve the model's performance in concept forgetting. Finally, exploring the integration of the Forget-Me-Not model with other techniques in text-to-image generation could lead to more advanced and effective solutions for handling unwanted concepts.

References

[1] Jorge Agnese, Jonathan Herrera, Haicheng Tao, and Xingquan Zhu. A survey and taxonomy of adversarial neural networks for text-to-image synthesis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(4):e1345, 2020.

[2] Fengxiang Bie, Yibo Yang, Zhongzhu Zhou, Adam Ghanem, Minjia Zhang, Zhewei Yao, Xiaoxia Wu, Connor Holmes, Pareesa Golnari, David A Clifton, et al. Renaissance: A survey into ai text-to-image generation in the era of large model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[3] Shengwen Ding and Chenhui Hu. Survey on the convergence of machine learning and blockchain. In *Proceedings of SAI Intelligent Systems Conference*, pages 170–189. Springer, 2022.

[4] Kaiyuan Gao, Sunan He, Zhenyu He, Jiacheng Lin, QiZhi Pei, Jie Shao, and Wei Zhang. Examining user-friendly and open-sourced large gpt models: A survey on language, multimodal, and scientific gpt models. *arXiv preprint arXiv:2308.14149*, 2023.

[5] Yijun Liu, Feifei Dai, Xiaoyan Gu, Minghui Zhai, Bo Li, and Meiou Zhang. Domain-aware and co-adaptive feature transformation for domain adaption few-shot relation extraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5275–5285, 2024.

[6] David Oniani, Jordan Hilsman, Yifan Peng, Ronald K Poropatich, Jeremy C Pamplin, Gary L Legault, and Yanshan Wang. Adopting and expanding ethical principles for generative artificial intelligence from military to healthcare. *NPJ Digital Medicine*, 6(1):225, 2023.

[7] Minh Pham, Kelly O Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. In *The Twelfth International Conference on Learning Representations*, 2023.

[8] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

[9] Yiyuan Yang, Ming Jin, Haomin Wen, Chaoli Zhang, Yuxuan Liang, Lintao Ma, Yi Wang, Chenghao Liu, Bin Yang, Zenglin Xu, et al. A survey on diffusion models for time series and spatio-temporal data. *arXiv preprint arXiv:2404.18886*, 2024.

[10] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017.