Robust Concept Erasure Using Task Vectors

Minh Pham, Kelly O. Marshall, Chinmay Hegde, and Niv Cohen
New York Univeristy
{mp5847,km3888,chinmay.h,nc3468}@nyu.edu

Abstract

With the rapid growth of text-to-image models, a variety of techniques have been suggested to prevent undesirable image generations. Yet, these methods often only protect against specific user prompts and have been shown to allow unsafe generations with other inputs. Here we focus on unconditionally erasing a concept from a text-to-image model rather than conditioning the erasure on the user's prompt. We first show that compared to input-dependent erasure methods, concept erasure that uses Task Vectors (TV) is more robust to unexpected user inputs, not seen during training. However, TV-based erasure can also affect the core performance of the edited model, particularly when the required edit strength is unknown. To this end, we propose a method called Diverse Inversion, which we use to estimate the required strength of the TV edit. Diverse Inversion finds within the model input space a large set of word embeddings, each of which induces the generation of the target concept. The learned set of word embeddings can be used to select the editing strength of a TV so that the model is more robust against unforeseen adversaries.

1. Introduction

The capacity of text-to-image (T2I) generative models to produce high-quality images has improved significantly over time. Consequently, growing concerns surround their potential for generating undesirable content. Such concerns include: ability to "deepfake" images of real people; ability to synthesize copyrighted materials; and production of Not-Safe-For-Work (NSFW) content. A direct approach to mitigate these would be to perform data filtering, i.e., removing all images depicting undesired concepts from the model's training set. However, automatically web-scraped, massive datasets are extremely hard to filter, and imperfect filtering often compromises the safety or the legal compliance of the resulting generative models (we refer to all kinds of undesirable generations as 'unsafe'). Additionally, even if filtering were feasible, retraining already existing models from scratch is often impractical due to high costs.

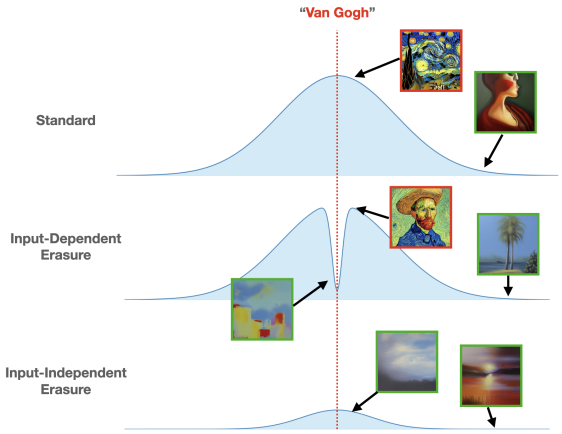

Figure 1. ***Input-independent vs. Input-dependent concept erasure.*** *Illustration of the probability distribution to generate the target concept "Van Gogh" across the input space. Images featuring the "Van Gogh" concept are framed in red, other images are framed in green. Input-dependent concept erasure leaves high probability areas of generating the target concept, while input-independent erasure methods erase the target concept across the entire input space.* ***(Top)*** *In generative T2I models, the probability of generating a specific concept is high for prompt embeddings close to the concept name, but high generation probability is possible also for prompts embedding in a significant distance from it.* ***(Middle)*** *Input-dependent concept-erasure attenuates the generation probability within a small environment of the given prompt but leaves a high probability of generating the erased concept further away from the prompt embedding.* ***(Bottom)*** *Input-independent erasure attenuates the probability of generating the target concept more consistently across the input space.*

Several recently proposed methods claim to "sanitize" unsafe concepts from T2I generative models [3, 6–8, 10, 19, 25]. Yet, when evaluated on unexpected inputs, these methods exhibit significant vulnerabilities [16, 23]. More specifically, most existing model sanitization approaches excel at averting the production of unsafe content, *conditioned on a specific* input prompt, or sequence of tokens.

In this paper, we aim to eliminate unsafe concepts from T2I models. Our core idea is based on a recently emergent

technique known as *Task Vectors* [9]. At a high level, a task vector (TV) represents a displacement in the model's weight space that is a result of fine-tuning; [9] shows that TVs can be flexibly used via arithmetic operations to enable editing of large models. Crucially, TV-based editing is independent of any specific user input, and therefore we showcase its ability to supply *unconditional* safety to T2I models. To apply TV-based concept erasure, we first fine-tune the model to generate a specific concept or style, and refer to the obtained weight difference as our TV. Next, we subtract the TV (possibly multiplied by a scalar α) from the original model, thereby erasing the unsafe concept.

Following the strong performance of TV edits for unconditional concept erasure on toy models, we investigate their application to large T2I models. Namely, we wish to apply TV edits while optimizing the trade-off between the erasure of unsafe concepts and the preservation of the model functionality.

Summary of our contributions. **(i)** Demonstrating that the vulnerability of current concept erasure methods is caused by their dependence on specific input prompts (Sec. 3) **(ii)** Using Diverse Inversion, we adjust editing strength based on a diverse set of dense prompts corresponding to a target concept to enhance the model's robustness against unforeseen adversaries (Sec. 4). **(iii)** Demonstrating TV-based editing as an efficient method for input-independent concept erasure (Sec. 5)

2. Related Work

Concept-Erasure on T2I Models. Recently, several strategies have been developed to prevent generative models from producing undesirable images. Negative Prompt (NP) [3] and Safe Latent Diffusion (SLD) [19] suggest modifying the inference process to divert the final output from undesired concepts. Other approaches employ classifiers to alter the output [1, 4, 18]. Since inference guiding methods can be evaded with sufficient access to model parameters [22], subsequent works including Erased Stable Diffusion (ESD) [6], Selective Amnesia (SA) [8], Forget-Me-Not (FMN) [25], Ablating Concepts (AC) [10], and Unified Concept Editing (UCE) [7] advocate for fine-tuning Stable Diffusion model weights.

Jailbraking T2I Models. As current concept erasure methods for T2I models are often reliant on protecting against specific user inputs, adversarial methods find other inputs that can induce unsafe generations. In particular, Tsai *et al*. [23] uses a CLIP text encoder to construct a concept vector; a vector in embedding space representing the unsafe content. It then uses a genetic algorithm [21] to find hard prompts that produce the concept vector in the embedding space. Additionally, Pham *et al*. [16] propose Concept Inversion, which is a method based on Textual Inversion [5] to search for word embeddings that circumvent concept erasure methods. Textual Inversion [5] learns to capture the user-provided concept by representing it through new "words" in the embedding space of a frozen T2I model without changing the model weights. In particular, the authors designate a placeholder string, c_*, to represent the new concept the user wishes to learn. They replace the vector associated with the tokenized string with a learned embedding v_*, in essence "injecting" the concept into the model vocabulary. The technique is referred to as Textual Inversion, whose optimization procedure is similar to [19] but we learn only the word embeddings instead.

Task Vectors and Parameter Space Interpolations. Although neural networks are inherently non-linear, Matena & Raffel [15] discovered that averaging the weights of multiple models, fine-tuned on different tasks from the same starting point can result in a model with high accuracy on all the fine-tuning tasks. Li *et al*. [13] observed similar outcomes when averaging the parameters of language models fine-tuned across various domains. Wortsman *et al*. [24] found that averaging the weights of models fine-tuned on multiple tasks can improve accuracy on a new downstream task without additional training. Interestingly, the weight difference learned during fine tuning can also be learned on one task and transferred to another to achieve a similar function. Like a vector, it can also be multiplied by a (possibly negative) scalar, and often conveys an appropriate meaning to the model function. Ilharco *et al*. [9] first compute a Task Vector (TV) as $\tau = \theta_{ft} - \theta_{pre}$, where θ_{pre} is the pre-trained model and θ_{ft} is the model fine-tuned on a selected set of tasks. Subtracting the TV, scaled by a constant α, from the pre-trained weights θ_{pre} will make the model perform worse on the selected tasks for which the fine-tuning process was done. On the other hand, adding a scaled TV will improve the model's performance on the same tasks. Ilharco *et al*. [9] show that Task Vectors, scaled by $\alpha \in [0, 1]$, can be applied to CLIP classifiers and LLMs to alter their behavior. In this work, we show that Task Vectors can also be applied to text-to-image diffusion models (in particular, the UNet module in Stable Diffusion) to perform concept erasure.

3. Motivating Analysis

While it is already known that common concept erasure methods may be circumvented by prompts not seen during the erasure process [16, 23], we find that they often filter only a small neighborhood around the embedding used for training (Fig. 2). Therefore, while existing methods are effective in blocking expected prompts, they are less robust to unexpected ones Fig. 1.

While evaluating the safety of a model independently from an input prompt is impractical, we can demonstrate input-independant safety on toy models. We hypothesize that prompt-independent concept erasure methods such as

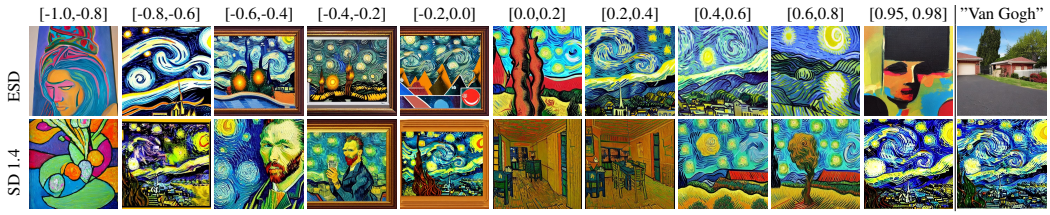

Figure 2. **Concept erasure methods often filter out only a tiny volume in input space**. Top row: Erased Stable Diffusion (ESD) with the Van Gogh" concept erased; bottom row: SD 1.4. We plot generations using adversarially optimized prompt embeddings at different cosine similarities to the Van Gogh" prompt embedding. Values in square brackets represent cosine similarities, ordered from left (far from concept) to right (closer to concept). ESD continues to produce "Van Gogh" concepts when the input prompt is far from the original concept name.

TV edits may provide better unconditional safety on a toy diffusion model. We test this hypothesis on a toy model we trained with dense "prompt" space of dimensions $d = 8$.

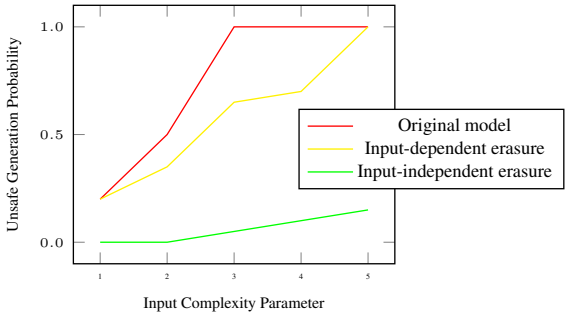

Figure 3. **TV-based concept-erasure provides better unconditional safety.** We plot the probability of unsafe generation with the most successful adversarial prompt from different exhaustive search resolutions. While input-dependent (finetune-based) concept erasure focuses on protecting against specific prompts, other prompts still produce unsafe generations with high probability. Input-independent (TV-based) erasure reduces the probability of unsafe generations compared to both the original and input-independent models across different complexity classes.

We trained our model to generate images from the MNIST [11] dataset. We apply three different concept-erasure models to erase the MNIST digit 0 from our diffusion model: (i) *Input dependent concept-erasure*: We fine-tune the model to produce the remaining 9 digits when given 0 as conditional input. (ii) *Input-independent concept-erasure:* We utilize a TV edit for input-independent concept-erasure [9]. We fine-tune our model to generate only the target concept, and then subtract the model weight change achieved by the fine-tuning process from the original model. (iii) *Original model:* We also evaluate the original model, without concept erasure. For all models, we use a pre-trained classifier to automatically evaluate whether the target concept was indeed generated.

To validate each model's safety in an input-independent manner, we perform an exhaustive search of the input space of our toy model at different resolutions. We perform our search in different resolutions, indexed by the number of examined values. We examine L possible values in each of the $d = 8$ dimensions (totalling L^d possible inputs). We refer to the different resolutions as our complexity parameter [12]. We can see in Fig. 3 that the TV edit provides a much better unconditional safety L_{uncod} input-independant guarantee.

4. Diverse Inversion: A New Method for Erasure Using Task Vectors

Having established TV-based editing as a method capable of improving the unconditional safety of T2I models, we now focus on applying this technique to larger models. Namely, we wish to erase unsafe concepts from large diffusion models while otherwise retaining their text-to-image capabilities. Measuring the degree of preservation of the desired text-to-image capabilities can be done directly, since this typically involves expected user inputs and outputs. However, anticipating the model's reaction to adversarial prompts *unknown* at the time of edit can be very challenging.

To estimate how well the model is protected against unexpected inputs, we would like to observe its outputs for a diverse array of adversarial prompts. We cannot inspect all the input prompts of a given length as we did for the toy model, due to the very large number of possible prompts. To this end, we create a diverse safety validation set composed of diverse input tokens that can all generate unsafe content. We note that a real-life adversary chooses their prompt after TV-based concept erasure has been applied, and not before it. Yet, the fact that an adversarial prompt often transfers well between erased and original (un-erased) models [16, 26] motivates us to rely on a large set of diverse adversarial prompts optimized for the original method.

4.1. Diverse Inversion

As we discuss in Sec. 3, concept erasure methods can provide a false sense of security by performing "input-

filtering". This suggests that additional inputs are needed to better evaluate concept erasure methods. We would like to have a diverse set on inputs, evaluating the concept erasure capability independently from any specific adversarial prompt. Hence, we learn multiple word embeddings simultaneously using Concept Inversion [16], which will then be used to measure erasure effectiveness of the edited model.

5. Experiments

Experimental setup To assess the content of the generated images, we use CLIP ViT-B/32 [17] pre-trained on LAION-2B [20]. Motivated by our experiment in Sec. 3, we propose to use a metric known as *Erasure Score* (ES) to validate the robustness of the edited Stable Diffusion model to many different attack prompts, defined as follows. After obtaining word embeddings via Diverse Inversion, we generate N images from the Stable Diffusion model using the learned embeddings and the concept name. Erasure score is defined as the maximum (calculated over the N images) CLIP similarity between the generated images and the concept name. A lower Erasure Score indicates more robustness against adversarial inputs. Our results on robustness to different adversarial methods are demonstrated qualitatively in Figs. 4 and 5.

Results We demonstrate in Figs. 4 and 5 that our method provides robustness to current adversarial methods applied after the concept erasure edit. In the second row in each of the sub-figures of Fig. 4 we show that for certain values of the edit strength α, the Stable Diffusion model manages to suppress the generation of the targeted concept when explicitly prompted with the same concept name. However, when Concept Inversion [16] is applied, we can still recover the erased concept. When α is increased we obtain more robust erasure. Moreover, as the larger α value makes the edited model more resistant to Concept Inversion, we also observe a decrease in Erasure Score (ES). This suggests that we can use the Erasure Score to guide us in selecting an appropriate edit strength, α, to make the model more robust against adversarial inputs. We also test our edited models against hard prompts obtained from the Ring-A-Bell method [23]. Fig. 5 shows that the adversarial prompts manage to fully circumvent 7 concept erasure methods but are unable to recover the erased targeted concepts using TV.

6. Limitations

Guarantees for erasure. An inherent weakness of any erasure method is the inability to evaluate them in advance against yet unknown future adversarial methods [2].

TV-based erasure. Our suggested method is reliant on TV techniques. Yet, the parameter space of neural networks is far from being completely understood [14]. The exact cases where TV-based erasure can work or fail are not clear yet.

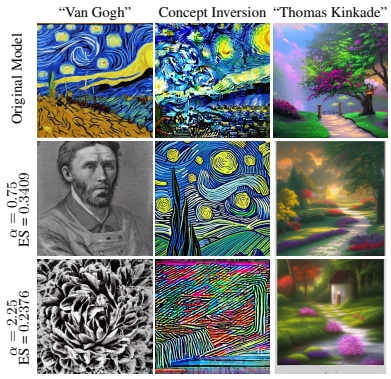

Figure 4. **TV-based concept erasure robustness to Concept Inversion.** We display three model variants (by row): the original model and two models with the targeted concept removed using Task Vectors of different magnitudes. For the "Van Gogh" concept, Stable Diffusion cannot generate images resembling Van Gogh's art style (even through Concept Inversion [16]) when erased using TV with sufficient magnitude. The 3^{rd} column shows that TV preserves model performance on unrelated concepts.

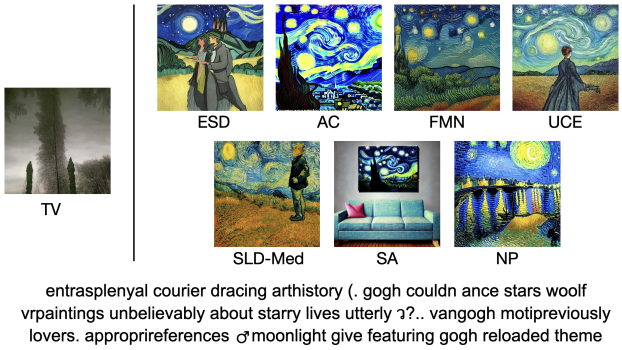

Figure 5. ***Generated images with the Ring-A-Bell [23] prompt for the concept "Van Gogh".*** *Adversarial prompt obtained from the "Ring-A-Bell" paper (bottom of the image) can circumvent 7 erasure methods, but not our suggested TV erasure procedure.*

The application of Task Vectors for more fine-grained, or coarse-grained concepts, is yet to be explored.

7. Conclusions

We propose adapting Task Vectors (TV), a recently proposed technique for model editing, for erasing concepts from generative models. On a range of test cases, we demonstrate how TVs can be used to sanitize undesirable concepts from text-to-image models in a way that is independent of specific user prompts. This independence distinguishes it from existing methods in the literature and facilitates its robustness.

References

[1] Stability AI. Stable diffusion 2.0 release, 2022. Jul 9, 2023. 2

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016. 4

[3] AUTOMATIC1111. Negative prompt, 2022. 1, 2

[4] Praneeth Bedapudi. Nudenet: Neural nets for nudity detection and censoring, 2022. 2

[5] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023. 2

[6] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *International Conference on Computer Vision*, 2023. 1, 2

[7] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzynska, and David Bau. Unified concept editing in diffusion models. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023. 2

[8] Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. In *Advances in Neural Information Processing Systems*, 2023. 1, 2

[9] Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023. 2, 3

[10] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *International Conference on Computer Vision*, 2023. 1, 2

[11] Yann LeCun, Corinna Cortes, and Christopher J.C. Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010. 3

[12] Ming Li, Paul Vitányi, et al. *An introduction to Kolmogorov complexity and its applications*. Springer, 2008. 3

[13] Margaret Li, Suchin Gururangan, Tim Dettmers, Mike Lewis, Tim Althoff, Noah A. Smith, and Luke Zettlemoyer. Branch-train-merge: Embarrassingly parallel training of expert language models. *CoRR*, abs/2208.03306, 2022. 2

[14] Chao Ma, Stephan Wojtowytsch, Lei Wu, et al. Towards a mathematical understanding of neural network-based machine learning: what we know and what we don't. *arXiv preprint arXiv:2009.10713*, 2020. 4

[15] Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems*, 2022. 2

[16] Minh Pham, Kelly O. Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. In *International Conference on Learning Representations*, 2024. 1, 2, 3, 4

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 4

[18] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. In *Advances in Neural Information Processing Systems Workshop*, 2022. 2

[19] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2

[20] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, 2022. 4

[21] S. N. Sivanandam and S. N. Deepa. *Introduction to genetic algorithms*. Springer, 2008. 2

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

[23] Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? In *International Conference on Learning Representations*, 2024. 1, 2, 4

[24] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, 2022. 2

[25] Eric J. Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. *CoRR*, abs/2303.17591, 2023. 1, 2

[26] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *CoRR*, abs/2307.15043, 2023. 3