Erasing Concepts from Diffusion Models with Free Hand Sketches

Venkatesh Tata
Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, United Kingdom
venkatesh.tata@surrey.ac.uk

(a) Apple Erased (b) Flight Erased

Figure 1: Results of concept erasure using the Concept Eraser ControlNet (CEC) model. (a) shows generated images when the concept of 'Apple' has been erased from the model, resulting in outputs where the apple is absent while other objects remain unaffected. (b) shows generated images when the concept of 'Airplane' is erased, where the model generates scenes devoid of airplanes while preserving the context of the sketches. This demonstrates the CEC model's ability to selectively remove specific concepts from diffusion models based on input sketches.

Abstract

This paper introduces a novel approach to enhancing content safety in generative models by erasing specific concepts from pre-trained diffusion models, particularly focusing on sketch-based image generation. Leveraging ControlNet, a state-of-the-art model for sketch-to-image tasks, we propose a method that modifies the internal weights of the model to eliminate undesirable content such as explicit or violent imagery. Our approach departs from traditional methods like dataset filtering mechanisms [1], post-generation filtering [2] and inference guiding [3], aiming to permanently remove unwanted concepts from the model's latent space, ensuring robustness against tampering or varied inputs. The method introduces the Classification Score, a new evaluation metric to evaluate if a CEM (Concept Erased Model), which in our case is the ControlNet from which target concept has been erased. We use ResNet50 pre-trained on ImageNet-1K to verify the success of concept erasure by analyzing top-1 predictions on the model's outputs. This ensures that undesired concepts are not only visually absent but are no longer encoded in the model's latent space. Through extensive experimentation, we identified key layers where concepts are encoded and selectively modified them to erase unwanted content while preserving the model's ability to generate other valid outputs. We demonstrate the robustness of this erasure across varied inputs, showcasing a reliable solution for enhancing content control in generative models. We release our code as open-source, available at [https://github.com/venkateshtata/Steering-Diffusion/tree/main](https://github.com/venkateshtata/Steering-

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Diffusion/tree/main), to encourage further research and facilitate community contributions toward enhancing content safety in generative diffusion models.

1 Introduction

In recent years, diffusion models have gained significant traction in the field of generative artificial intelligence, particularly in image synthesis. These models excel at generating high-quality images based on input conditions, such as text prompts or sketches. However, as these models become increasingly sophisticated, concerns have emerged around their potential to generate harmful or undesirable content, including explicit, violent, or otherwise inappropriate imagery. This issue poses ethical challenges for the deployment of AI systems in public-facing applications, where ensuring the safety and appropriateness of generated content is crucial.

Traditionally, methods to mitigate this problem involve filtering mechanisms [2] that block unwanted outputs based on the analysis of input conditions, such as textual descriptions. However, these methods often remain vulnerable to circumvention and tampering, as the core model retains its ability to generate undesirable content. Moreover, diffusion models can sometimes generate random or irrelevant content, resulting in lower-quality outputs that detract from their effectiveness in applications requiring precise and meaningful image generation.

This dissertation addresses these challenges by exploring a more robust approach: the complete erasure of undesirable concepts from the model at the weights level. The project specifically leverages ControlNet [4], a leading diffusion model for sketch-based image generation, to demonstrate how concept erasure can be achieved by integrating both sketch and text inputs. By targeting the core model, this approach ensures that once the unwanted concepts are removed, the model is no longer capable of generating such content, even if tampered with.

1.1 Background and Context

Diffusion models have become a cornerstone in the field of generative artificial intelligence, particularly for generating high-quality images from various input modalities such as text and sketches. These models, while powerful, pose risks due to their ability to generate unwanted content, such as explicit or violent imagery, which raises significant ethical concerns regarding their deployment in real-world applications. ControlNet, a diffusion model tailored for sketch-based image generation, stands out as a state-of-the-art approach. However, its ability to produce inappropriate or random content necessitates solutions that go beyond surface-level filtering.

This paper addresses these ethical challenges by proposing a novel method that erases specific concepts from diffusion models at the core weights level, thereby preventing the generation of unwanted imagery. Unlike traditional filtering methods, which block the input from producing certain outputs, this method ensures that the model itself is incapable of generating undesirable content.

The work builds upon ControlNet's sketch-based generation capabilities, employing sketches and text as dual conditions to target and erase specific concepts, such as "airplane," "cat," "apple," and others more effectively. Extensive experiments, including those presented in the below sections, highlight the robustness of concept erasure, ensuring that the model no longer generates the erased concepts even when different sketches of the same concept are provided.

2 Related Work

2.1 Latent Space Manipulation in Generative Models

Recent research has aimed at improving control over latent space manipulation in generative models. [5] introduced a combination of an auxiliary map and Conjugate Gradient to enhance precision, while [6] focused on distribution-preserving transport maps to address the mismatch between latent space operations and prior distributions. [7] developed a matrix subspace projection method for disentangling attribute information in autoencoders, enabling more precise manipulation of attributes without affecting others. [8] found that simple linear mappings between latent spaces can preserve key characteristics, enhancing cross-model compatibility. Collectively, these advances improve control,

disentanglement, and distribution alignment in latent space operations, benefiting domains like image and text generation.

2.1.1 Overview of Latent Space in Generative Models

Latent spaces are central to deep generative models, capturing relationships in data and enabling complex manipulations. The geometric structure of these spaces impacts model performance and flexibility. [9] showed that a properly dimensioned latent space, with more dimensions than the data's modes, improves model accuracy and performance. They introduced a truncation method to impose a cluster structure in the latent space, enhancing adherence to target distributions in GANs.

Nonlinear mapping between latent and input spaces can distort operations like distance measurement and interpolation. [10] proposed a stochastic Riemannian metric to correct these distortions, leading to improved sample quality and better variance estimates in generative models. These insights emphasize the importance of latent space geometry in improving model performance.

2.1.2 Techniques for Latent Space Navigation

Recent methods enable creative navigation of latent spaces. [11] introduced a tool allowing users to design trajectories through latent spaces, facilitating time-based media like videos. [12] developed a controller module for smooth traversal in Variational Autoencoders (VAEs), aiding in video editing and animation.

Schwettmann et al. (2020) [29] created "Latent Compass," mapping human-interpretable features to latent space directions for more intuitive image manipulations. This method enhances creativity by allowing real-time interaction with latent spaces.

2.1.3 Challenges in Latent Space Manipulation

Latent space manipulation faces challenges, particularly in disentanglement and dimensionality. [5] addressed these by combining an auxiliary map with Conjugate Gradient, improving precision and control. [13] tackled feature entanglement by compressing the latent space, leading to more efficient exploration and manipulation. [14] introduced Surrogate Gradient Fields (SGF) for multidimensional condition manipulation, enabling more sophisticated image modifications, such as key points and captions.

In shape editing, [15] introduced a method that disentangles latent sub-spaces into style variables and control points, allowing intuitive manipulation of object shapes. These approaches advance the precision and usability of latent space manipulation, enhancing the ability to make complex modifications in generative models.

2.2 Concept Erasing in Diffusion Models

2.2.1 Concept Erasure Techniques in Diffusion Models

Recent research has focused on erasing specific concepts from text-to-image diffusion models to prevent the generation of undesirable content, such as explicit imagery, copyrighted material, or biased concepts. The challenge is achieving robust erasure—blocking concept reappearance under paraphrased prompts—while maintaining the model's ability to generate unrelated content.

Receler (Reliable Concept Erasing via Lightweight Erasers) by [16] proposes a lightweight method using adversarial prompt learning and concept-localized regularization. This approach effectively removes concepts with minimal modifications to the model, ensuring that erased concepts do not resurface under paraphrased prompts. Receler excels in maintaining the model's capability to generate non-erased content, making it a precise tool for fine-grained content control.

[17] introduced a fine-tuning method using negative guidance to permanently erase specific visual concepts from diffusion models. This technique is particularly effective at removing unwanted styles such as explicit content or copyrighted material, ensuring that similar prompts do not regenerate the erased content. It surpasses methods like Safe Latent Diffusion, offering a more robust solution for concept removal without compromising unrelated outputs.

2.2.2 Advances in Scalable and Efficient Concept Erasure

[18] developed a fast and efficient few-shot unlearning method that updates the text encoder using only a few real images. This technique achieves rapid concept erasure in seconds, while ensuring the model continues to perform well on unrelated tasks. The implicit transition to related latent concepts results in seamless concept removal, making this method particularly useful in scenarios where speed and efficiency are crucial.

The latest advancement, MACE (Mass Concept Erasure), by [19], expands the scope of concept erasure by handling up to 100 concepts simultaneously. Using closed-form cross-attention refinement and LoRA finetuning, MACE balances generality and specificity, ensuring erased concepts, including objects, celebrities, and artistic styles, do not reappear. MACE excels in benchmarks, outperforming previous methods while preserving the model's ability to generate diverse outputs.

These methods represent significant progress in creating safer and more controlled generative models, addressing ethical concerns while retaining creative potential across various applications.

3 Methodology

3.1 Dataset and Pre-Processing Input Condition

For our experimentation and evaluation, we utilized the Sketchy Database, a comprehensive dataset specifically designed for sketch-based image generation and retrieval tasks. The Sketchy Database contains over 75,000 unique sketches and corresponding images across 125 object categories, ranging from everyday items like "apple" and "cat" to more abstract entities such as "airplane" and "fish." Each sketch in the dataset is manually drawn, offering a diverse array of visual representations that vary in style, detail, and abstraction level.

The selection of the Sketchy Database was motivated by its broad coverage of object categories and its utility in testing the effectiveness of concept erasure in diffusion models. The dataset's high-quality and category-diverse sketches make it well-suited for evaluating the generalizability and robustness of the CEM. Since the core of our methodology focuses on sketch-based conditions, the use of a dataset rich in sketches provides an ideal environment for evaluating the model's ability to learn, recognize, and erase specific visual concepts across multiple instances. ControlNet leverages a robust pre-processing pipeline to convert raw sketches into structured inputs that guide the image generation process. While ControlNet supports diverse input conditions such as depth maps and human poses, this work focuses on using sketches to exert precise control over the diffusion process.

3.1.1 Sketch Normalization, Edge Detection, and Structuring Input Conditions for ControlNet

The first step in processing sketches involves using Holistically Nested Edge Detection (HED) to transform hand-drawn or digital sketches into refined edge maps that capture essential contours while reducing noise. These edge maps are resized to 512x512 pixels for model compatibility, and optional Gaussian blurring is applied to smooth irregularities. The edge maps are then normalized (0-1 range) to facilitate consistent model inference and serve as conditioning inputs for ControlNet. ControlNet combines these structural maps with text prompts, guiding output generation by aligning the sketch's structure with semantic details. This process retains the pre-trained diffusion model's parameters while adding layers for controlled output, integrating both structural and textual cues.

3.2 Training ControlNet for Concept Erasure

This section describes the training process for ControlNet to erase specific concepts using multi-modal inputs (text and sketches) as shown in Figure-2. This approach extends Gandikota's work [17] by incorporating both input types, enabling selective concept removal from a pre-trained diffusion model.

3.2.1 Objective of Concept Erasure

The goal is to teach ControlNet to erase specified concepts (e.g., "airplane") based on text and sketch inputs. The model must generalize erasure across variations of the concept while preserving other

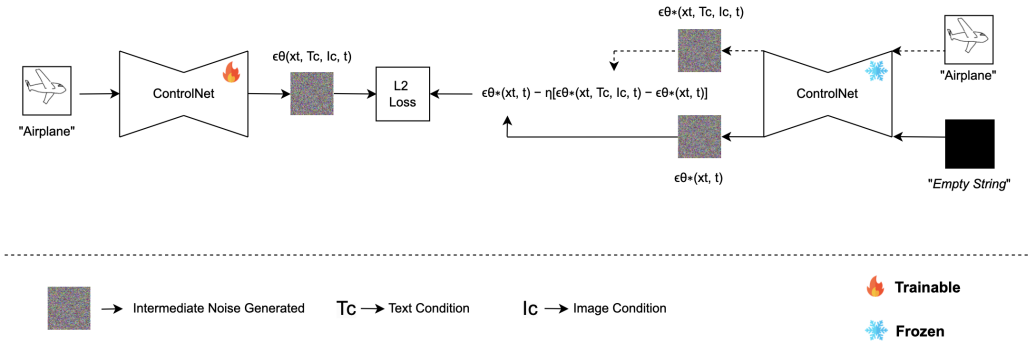

Figure 2: Overview of the concept erasure process in diffusion models using ControlNet. The left side depicts an initial sketch input of an 'Airplane,' which is processed through ControlNet conditioned on both text (Tc) and image (Ic) inputs. The trainable components adjust the model's output to systematically erase the specified concept, guided by an L2 loss function that minimizes the influence of the conditioned concept on the generated noise. The right side shows the updated output after concept erasure, with the ControlNet model producing an output devoid of the 'Airplane' concept when conditioned with an 'Empty String' input, illustrating successful concept removal

elements. ControlNet modifies trainable layers to erase targeted concepts while the base diffusion model remains fixed. The challenge is to ensure that erasure is consistent across multiple variations of the concept (styles, orientations) without affecting other image features.

3.2.2 Training Setup and Loss Function Design

Training employs a loss function based on denoising score matching, guiding the model to forget the unwanted concept. The loss measures the difference between the trainable ControlNet's output conditioned on text and sketches and the frozen ControlNet's unconditioned output.

The loss function is:

$$L = \|\epsilon_\theta(x_t, T_c, I_c, t) - (\epsilon^*_\theta(x_t, t) - \eta\left[\epsilon^*_\theta(x_t, T_c, I_c, t) - \epsilon^*_\theta(x_t, t)\right])\|^2$$

Where:

- $\epsilon_\theta(x_t, T_c, I_c, t)$ represents the output of the trainable ControlNet conditioned on noisy input x_t, text T_c, and sketch I_c.
- $\epsilon^*_\theta(x_t, t)$ is the unconditioned output from the frozen ControlNet.
- η controls the strength of erasure by scaling the corrective adjustment $[\epsilon^*_\theta(x_t, T_c, I_c, t) - \epsilon^*_\theta(x_t, t)]$.

The objective is to minimize the influence of the conditioned concept on the output, thereby erasing the concept from the generated images.

4 Experiments

This section provides a comprehensive evaluation of the Concept Erased Model (CEM) following its fine-tuning, with a focus on assessing the effectiveness of concept erasure, robustness across various input conditions, classifier impact analysis, and the behavior of the loss function during training.

The model was fine-tuned on an NVIDIA A6000 GPU with 48 GB of RAM, ensuring efficient processing and memory utilization. Extensive experimentation determined that 500 iterations were sufficient to achieve complete erasure of a specified target concept while maintaining the model's capacity to generate non-target classes accurately. Through parameter optimization, we identified the key hyperparameters that yielded optimal performance: a learning rate set at 1×10^{-5} and a

negative guidance coefficient of 1. These parameters facilitated effective concept erasure without compromising the generation of unrelated content.

Additionally, all images were generated at a resolution of 512x512 pixels to maintain consistency across experiments and to ensure that the erasure was effectively applied at a sufficiently detailed level. The optimal number of steps for the Denoising Diffusion Implicit Models (DDIM) process was determined to be 50, balancing the trade-off between computational efficiency and high-quality output. This setup provided the most robust results in terms of the model's ability to remove unwanted concepts while retaining visual fidelity in the generated outputs.

4.1 Erased Class vs Un-erased Class

The experiment tested the model's ability to erase specific classes like "airplane" and "apple" while retaining the generation of un-erased classes like "banana" and "cat" as shown in Figure-1 After erasure, sketches of erased objects failed to generate relevant imagery, while un-erased classes were unaffected, demonstrating effective, selective erasure.

4.2 Erasure Robustness

(a) Cat Erased CEM Output (b) Fish

Figure 3: Sketch-to-Image Generation with Concept Erasure: The top row displays various free-hand sketches of a cat. The bottom row shows the resulting images generated by a diffusion model conditioned on these sketches after applying the Concept Eraser ControlNet (CEC). As demonstrated, the CEC model alters the output to obscure the representation of 'cat,' yielding abstract, distorted, or unrecognizable images that differ significantly from the original sketches. This illustrates the effectiveness of the concept erasure mechanism in preventing the model from generating the erased concept.

To evaluate the robustness of the Concept Erased Model (CEM), we assessed its ability to consistently remove specified concepts across a variety of sketches differing in style, pose, and orientation. The model was presented with diverse representations of the same concept, including abstract, detailed, simplistic, and exaggerated drawings, to test whether it could maintain concept erasure without generating recognizable forms of the erased concept. As shown in Figure 1, the CEM effectively obscures the erased classes (e.g., cat, airplane) regardless of sketch complexity or style, indicating its capacity to generalize and erase concepts consistently across varied visual inputs.

The model's robustness was further tested with challenging cases, including extreme poses, unconventional angles, and intricate details. While the CEM successfully prevented the reappearance of erased concepts in most instances, minor abstract traces occasionally emerged in highly detailed or contextually complex sketches, though these remained largely unrecognizable. This evaluation underscores the CEM's effectiveness in ensuring that erased concepts remain absent across diverse real-world scenarios, highlighting its potential for enhancing content control in generative diffusion models.

5 Classification Score for Evaluating Concept Erasure

In this work, we introduce the "Classification Score" as a novel metric to evaluate the effectiveness of concept erasure in CEMs, specifically within the ControlNet framework. The Classification Score leverages a pre-trained ResNet50 classifier (trained on ImageNet-1K) to determine whether an erased concept still persists in the outputs generated by the CEM. The key hypothesis is that a successful

concept erasure will prevent the classifier from identifying the erased class, even when presented with a sketch corresponding to that concept.

The evaluation process begins by inputting a sketch of the target concept into the Concept Erased Controlnet, which generates an image where the specified concept has been erased. This generated image is subsequently analyzed by the ResNet50 classifier. If the classifier's top prediction corresponds to the erased concept, it suggests that the erasure was unsuccessful. However, if the classifier consistently predicts unrelated classes, it indicates that the erasure has succeeded, demonstrating that the concept has been removed both visually and within the latent space of the model.

Table 1: Classification Scores (ResNet50 on ImageNet 1K) for CEC Model Outputs: The table reports the top-1 predictions of a pre-trained ResNet50 classifier for images generated by the concept erased controlnet model when sketches of various concepts (Cat, Airplane, Apple, Banana, Fish) are provided as input. Each row corresponds to an erased class and shows how the classifier's predictions differ across sketches of various concepts. Successful concept erasure is indicated when the classifier predicts classes unrelated to the erased concept.

Input Image:	**Cat**	**Airplane**	**Apple**	**Banana**	**Fish**
Erased Class:					
Cat	judge's robe	wing	mask	banana	coral reef
Airplane	Egyptian cat	wing	Apple	banana	coral reef
Apple	tabby cat	space shuttle	lens cover	banana	goldfish
Banana	fountain	airliner	Apple	lamp shade	coral reef
Fish	tabby cat	airliner	tray	banana	sea slug

The Classification Score serves as a quantitative measure of the CEM's success, confirming that the erased concept is thoroughly suppressed not only visually but also in its latent representations. This metric significantly contributes to enhancing content safety in generative models, ensuring that undesired content does not reappear after erasure. Moreover, this highlights the potential of using pre-trained classifiers as powerful tools for evaluating and controlling the behavior of generative models in future research endeavors.

6 Limitations

Despite the successful erasure of target concepts, the Concept Erased Model (CEM) exhibits some limitations that warrant further investigation. One of the primary concerns is the degradation of output quality for non-target classes. While the model effectively erases the specified concept, it often produces low-quality or distorted representations of non-erased classes. This degradation varies depending on the specific target concept being erased, suggesting that the erasure process may impact the model's generative capabilities beyond the intended scope.

Additionally, when the CEM is conditioned on sketches and textual prompts corresponding to the erased target class, the model often generates outputs characterized by noise and visual artifacts instead of coherent, contextually appropriate imagery. This behavior indicates a potential failure in redirecting the generative process to produce semantically meaningful alternatives once the target concept is removed. To address this, future work should focus on improving the model's capacity to generate high-quality, conceptually sound outputs even when faced with inputs corresponding to erased concepts, ensuring both robustness and fidelity in diverse input scenarios.

7 Conclusion

In conclusion, this research presents a novel method for concept erasure in diffusion models, specifically leveraging ControlNet in sketch-based image generation. By modifying the model at the weight level, this approach effectively erases specific concepts, ensuring they cannot be regenerated, even when prompted with related inputs. The method integrates both sketch and text inputs to guide the model during the erasure process, and its effectiveness was validated through comprehensive evaluation, including visual inspection and a newly introduced Classification Score. These evaluations demonstrated the method's robustness across a range of input conditions and perspectives.

Beyond the technical contributions, this work addresses significant ethical concerns in AI, offering a solution to prevent harmful or undesirable content generation directly at the model level. By ensuring the erasure of unwanted concepts, this method contributes to the development of safer, more reliable generative AI systems, highlighting its potential for broader applications in controlled AI deployment practices.

8 Future Work

While the research demonstrated successful concept erasure, there are several areas for future exploration. One key area is improving the robustness of concept erasure across extreme variations of input conditions. Expanding the range of sketches and styles used during training could further enhance the model's ability to generalize erasure across more diverse and abstract representations of the erased concept.

Another avenue for future work is the exploration of multi-modal erasure, where concepts are erased not only based on visual sketches but also on other modalities such as audio, video, or 3D representations. This would broaden the applicability of the method in more complex generative AI systems.

Additionally, while the Classification Score provided a reliable metric for evaluating the effectiveness of concept erasure, future research could develop more sophisticated evaluation metrics that incorporate human perceptual assessments or adversarial testing to ensure that erased concepts are not only absent in a classifier's predictions but are also imperceptible to human observers.

Finally, scaling this method to larger and more complex diffusion models, and applying it to real-world scenarios where content safety is paramount, such as in autonomous systems, media generation, or social platforms, would be a valuable extension of this research. Further studies could also investigate the ethical implications and ensure that the approach aligns with broader AI governance and safety standards.

References

[1] 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, 2022.

[2] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter, 2022.

[3] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models, 2023.

[4] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. 2 2023.

[5] Walid Messaoud, Rim Trabelsi, Adnane Cabani, and Fatma Abdelkefi. Conjugate gradient for latent space manipulation. In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence*, pages 50–57. SCITEPRESS - Science and Technology Publications, 2024.

[6] Eirikur Agustsson, Alexander Sage, Radu Timofte, and Luc Van Gool. Optimal transport maps for distribution preserving operations on latent spaces of generative models. 11 2017.

[7] Xiao Li, Chenghua Lin, Ruizhe Li, Chaozheng Wang, and Frank Guerin. Latent space factorisation and manipulation via matrix subspace projection. 7 2019.

[8] Andrea Asperti and Valerio Tonelli. Comparing the latent space of generative models. 7 2022.

[9] Thibaut Issenhuth, Ugo Tanielian, Jérémie Mary, and David Picard. Unveiling the latent space geometry of push-forward generative models. 7 2022.

[10] Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. 10 2017.

[11] Memo Akten, Rebecca Fiebrink, and Mick Grierson. Deep meditations: Controlled navigation of latent space. 2 2020.

[12] Yan Zuo, Gil Avraham, and Tom Drummond. Traversing latent space using decision ferns. 12 2018.

[13] Chien-Hung Lin, Yi-Lun Pan, and Ja-Ling Wu. Attribute-based facial image manipulation on latent space. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, 2021.

[14] Minjun Li, Yanghua Jin, and Huachun Zhu. Surrogate gradient field for latent space manipulation. 4 2021.

[15] Tim Elsner, Moritz Ibing, Victor Czech, Julius Nehring-Wirxel, and Leif Kobbelt. Intuitive shape editing in latent space. 11 2021.

[16] Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. 11 2023.

[17] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. 3 2023.

[18] Masane Fuchi and Tomohiro Takagi. Erasing concepts from text-to-image diffusion models with few-shot unlearning. 5 2024.

[19] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. 3 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction of the paper accurately reflect the main contributions, focusing on the erasure of specific concepts from pre-trained diffusion models using sketch-based image generation and ControlNet. The claims are well-supported by the technical content and experimental results presented in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a section on limitations that discusses the degradation of output quality for non-target classes and the occasional generation of noisy or distorted imagery when conditioned on erased concepts. These limitations are clearly articulated and aligned with the scope of the research.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper primarily focuses on applied research in diffusion models and does not include theoretical proofs or assumptions. The methodology and experiments are empirically driven, and there are no theorems or formal mathematical results requiring proof.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides sufficient details on the experimental setup, including datasets (Sketchy Database), training parameters (learning rate, iterations), and hardware used (NVIDIA A6000 GPU). This information is adequate to reproduce the main results and verify the claims made in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides open access to its code through a GitHub repository (https://github.com/venkateshtata/Steering-Diffusion/tree/main). This repository enables the research community to replicate the experiments, validate the results, and build upon the work, facilitating transparency and further advancements in the field of content control for diffusion models. Additionally, sufficient instructions are provided to ensure that others can faithfully reproduce the main experimental results presented in the paper.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental section provides comprehensive details, including dataset specifications, training parameters, pre-processing techniques (e.g., Holistically Nested Edge Detection for sketches), and the use of classification scores to evaluate concept erasure. This thorough documentation allows for a clear understanding of the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The paper presents experimental results with visual and qualitative analysis but does not report error bars, confidence intervals, or statistical tests for significance. This is a point of improvement, as providing statistical significance could strengthen the experimental findings.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper specifies the hardware used (NVIDIA A6000 GPU with 48 GB RAM) and provides information on the number of iterations (500), learning rates, and model steps used. This information helps to understand the computational resources required to reproduce the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper addresses ethical concerns by proposing a method to erase potentially harmful or undesirable content from generative models, aiming to contribute positively to AI safety and ethics. It adheres to responsible AI practices in the discussion of its methodology and impact.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader impacts section discusses both the positive and negative societal implications of the proposed concept erasure technique. The paper highlights its contribution to AI safety and content moderation while recognizing the need for further improvement to address visual quality degradation for non-erased classes.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not introduce new models or data that pose a high risk for misuse, such as large-scale language models or sensitive datasets. Therefore, no additional safeguards are discussed.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper uses the Sketchy Database and pre-trained ResNet50 on ImageNet-1K, which are publicly available. The original sources are properly credited, though the paper does not explicitly state the licenses for these assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release any new datasets, models, or code assets, so there is no documentation provided for new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects, making this question not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Since the paper does not involve any human subjects research, IRB approvals or similar are not required, making this question not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.