Style Unlearning: Leveraging Parameter-Efficient Modules for Efficient Style Removal

Linxi Xie*
New York University
New York, USA
lx2154@nyu.edu

Xinyue Sheng*
New York University
New York, USA
xs2334@nyu.edu

Yuhan Liu*
New York University
New York, USA
yl10379@nyu.edu

Zhuoran Chen*
New York University
New York, USA
zc2745@nyu.edu

GitHub: GitHub Repository

***Abstract*—Motivated by concerns that large-scale diffusion models can generate undesirable outputs, such as toxic content or copyrighted artistic styles, we focus on unlearning harmful or undesired styles while addressing copyright issues. We propose two efficient methods for style unlearning, leveraging Low-Rank Adaptation (LoRA) modules from the CivitAI platform. Unlike traditional approaches that require retraining and access to model weights, our methods—Negation of LoRA and Weights2Weights (w2w) unlearning—eliminate these inefficiencies. Through extensive experiments, we evaluate these methods in terms of preservation and effectiveness using CLIP and Vision-Language Models (VLMs). Results show that W2W achieves more consistent and robust style suppression compared to Negation of LoRA, while both methods maintain content fidelity. This work highlights the potential of parameter-efficient frameworks for targeted style unlearning, offering scalable solutions to enhance control in diffusion-based image generation.**

I. Introduction

The rapid advancement of diffusion models has revolutionized image generation, enabling the production of high-resolution, high-quality images across diverse styles. However, challenges remain, particularly concerning the generation of undesired or harmful styles, and the need to respect copyrights associated with specific artistic styles. Addressing these challenges requires methods to effectively unlearn certain styles, ensuring that the model avoids generating them regardless of the user's prompts. Moreover, it is crucial to maintain the model's ability to fulfill other prompt requirements after unlearning. Existing approaches to style unlearning typically involve fine-tuning pretrained models by modifying their weights with new objectives. While effective, these methods are computationally expensive, requiring access to the model's internal parameters and substantial retraining efforts. To overcome these inefficiencies, we propose novel methods utilizing existing LoRAs on the civitai platform, which eliminate the need for retraining or knowledge of model weights. Our framework introduces two efficient techniques: Negation of LoRA, which systematically suppresses undesired styles by applying a negative weight to LoRA modules, and Weights2Weights (W2W) unlearning, which extends this concept by leveraging a reduced-dimensionality subspace to identify and suppress style-specific attributes. Together, these methods provide a lightweight yet effective solution for targeted style unlearning, preserving the model's overall performance while enhancing its flexibility and interpretability. Our study will explore the following research questions:

RQ 1: How effective is Negation of LoRA in suppressing specific styles without affecting overall image content?

RQ 2: Can the Weights2Weights (W2W) framework identify meaningful subspaces for targeted style unlearning, and how does it compare to one-dimensional attribute negation in terms of preservation and effectiveness?

RQ 3: What are the trade-offs between preserving the original generative quality and successfully unlearning undesired styles when using the proposed methods?

II. Related Work

A. Machine Unlearning

Machine unlearning was first developed in Large Language Models(LLM) to mitigate harmful training data's influence on the model. Multiple methods were proposed to achieve this unlearning outcome. Yao et al. utilized Gradient Ascent and KL-divergence to achieve unlearning of the model [1]. Chen and Yang introduce lightweight unlearning layers to handle forgetting operations [2]. Wang et al. [3] propose a method to align the knowledge between the pre-trained model and fine-tuning model.

B. Unlearning Methods in Image Generation

Recent advancements in unlearning processes focus on two aspects: removing targeted attributes or styles while preserving other components. For toxic concept removal, Han et al. (2023) introduced Unified Concept Editing (UCE) [4] , which modifies cross-attention layers to target specific concepts. Similarly, Juwon Sao et al. [5] proposed the GUIDE framework to remove specific identities from generated images with minimal training data. For style removal, methods like ESD [6] edit pre-trained diffusion U-Net model weights, and tools like UNLEARNCANVAS [7] provide benchmarks for assessing style removal. While these approaches achieve notable results, they rely on modifying pre-trained model weights, highlighting the need for more efficient, lightweight methods.

[0*] Authors are listed in alphabetical order by their first names.

C. Parameter-efficient Methods

LoRA leverages low-rank adaptation to efficiently modify specific parameters by adding or subtracting task-specific vectors, optimizing parameter efficiency [8]. Building on this concept, Zhang et al. (2023) utilized Parameter-Efficient Modules (PEMs) to achieve machine unlearning. They trained PEMs on toxic data from the Civil Comments dataset and then negated them on the GPT-2 large language model, achieving substantial reductions in model toxicity with minimal impact on linguistic proficiency. Inspired by this success, we aim to investigate whether negating PEMs could yield similar results in the image domain.

Further expanding this idea, the paper Interpreting the Weight Space of Customized Diffusion Models introduced the term ”weights2weights” (W2W) [9], which represents a subspace of the weight space derived from a large collection of customized diffusion models. By identifying linear directions in this subspace corresponding to semantic edits, they demonstrated the ability to create new models with specific identities modified. This approach highlights a novel perspective on using PEMs, such as LoRA, for concept editing, offering potential applications in targeted concept removal.

III. Methods

A. Negation of LoRA

Low-Rank Adaptation (LoRA) [8] is a parameter-efficient fine-tuning technique originally developed for large language models. It factorizes the parameter update ΔW into two low-rank matrices A and B, thereby greatly reducing the number of trainable parameters. Formally, if W_0 denotes the original pretrained weights of the model and ΔW represents the LoRA-induced update, the adapted weights can be expressed as:

$$W = W_0 + \alpha \Delta W, \quad \Delta W = AB$$

where α is a scalar controlling the influence of the LoRA update. Integrating LoRA into the cross-attention layers of a U-Net-based diffusion model allows for efficient injection of specific styles or concepts into generated images without retraining the entire model.

While LoRA is typically employed with $\alpha > 0$ to enhance a given attribute, the same parameterization suggests that negating the weight ($\alpha < 0$) could systematically suppress it. We refer to this process as **Negation of LoRA**. If a particular LoRA module encodes an attribute direction in the model's parameter space (e.g., amplifying ”anime” style features), applying $\alpha < 0$ theoretically pushes the model's output away from that attribute. In essence, negating LoRA attempts to unlearn a given characteristic by moving the model's parameters in the opposite direction of the learned attribute vector.

However, as this negation represents a one-dimensional traversal along the attribute direction, it may not consistently yield a high-quality or semantically meaningful opposite style. The results presented in the IV.B section highlight the practical outcomes and limitations of this method.

B. Unlearning with weight to weight

We further consider a Weights2Weights (W2W) framework to extend beyond one-dimensional attribute negation to learn a more meaningful and controlled unlearning direction. Building on prior knowledge from identity editing in W2W [9], we adapt this concept for style-based unlearning.

To create a W2W space tailored for image styles, we curated a dataset of model weights, denoted as $D = \{\theta_1, \theta_2, \ldots, \theta_N\}$. Each weight θ_i was generated by flattening and concatenating the LoRA matrices of fine-tuned models, forming data points $\theta_i \in \mathbb{R}^d$, where each point represents a distinct image style. To reduce dimensionality and identify meaningful subspaces, we applied Principal Component Analysis (PCA) on the dataset, retaining the top m principal components. This process established a basis of vectors $\{w_1, w_2, \ldots, w_m\}$, enabling us to represent each weight as a linear combination within this lower-dimensional subspace.

To achieve targeted style unlearning, we sought a direction $v \in \mathbb{R}^d$ in the weight space that separates styles effectively. Using binary labels obtained for each model (e.g., anime/realistic), we trained linear classifiers with model weights as input features. The hyperplane determined by the classifier separates models based on style attributes, and the normal vector v to this hyperplane serves as the traversal direction. Given a model weight θ representing an image style, unlearning is achieved by moving orthogonally along the direction v. The edited weights are calculated as:

$$\theta_{\text{edit}} = \theta + \alpha v$$

where α is a scalar controlling the strength of the unlearning operation. This adjustment modifies the model weights to suppress the target style while preserving other features. A single unlearning operation in the W2W space results in a new model that ideally no longer generates images characteristic of the unlearned style. This framework ensures a controlled and interpretable unlearning process, leveraging the linear properties of the W2W space to navigate and edit style-specific representations effectively.

C. Evaluation Methods

We evaluate the effects of negation from two perspectives: (1) Preservation: We calculate the CLIP score between the generated images and a content-only prompt, excluding any style elements, to assess how well the core content is retained across outputs from the base model and its variations with LoRAs. (2) Effectiveness: We use two complementary methods to evaluate stylistic alignment. First, we compute the CLIP score between the style-specific prompt and the generated image. Second, we prompt a Vision-Language Model (VLM) to analyze the image and provide a rating based on how closely it aligns with the style description. A more stable preservation score reflects better content preservation, while lower effectiveness scores in VLM and CLIP indicate stronger elimination of style

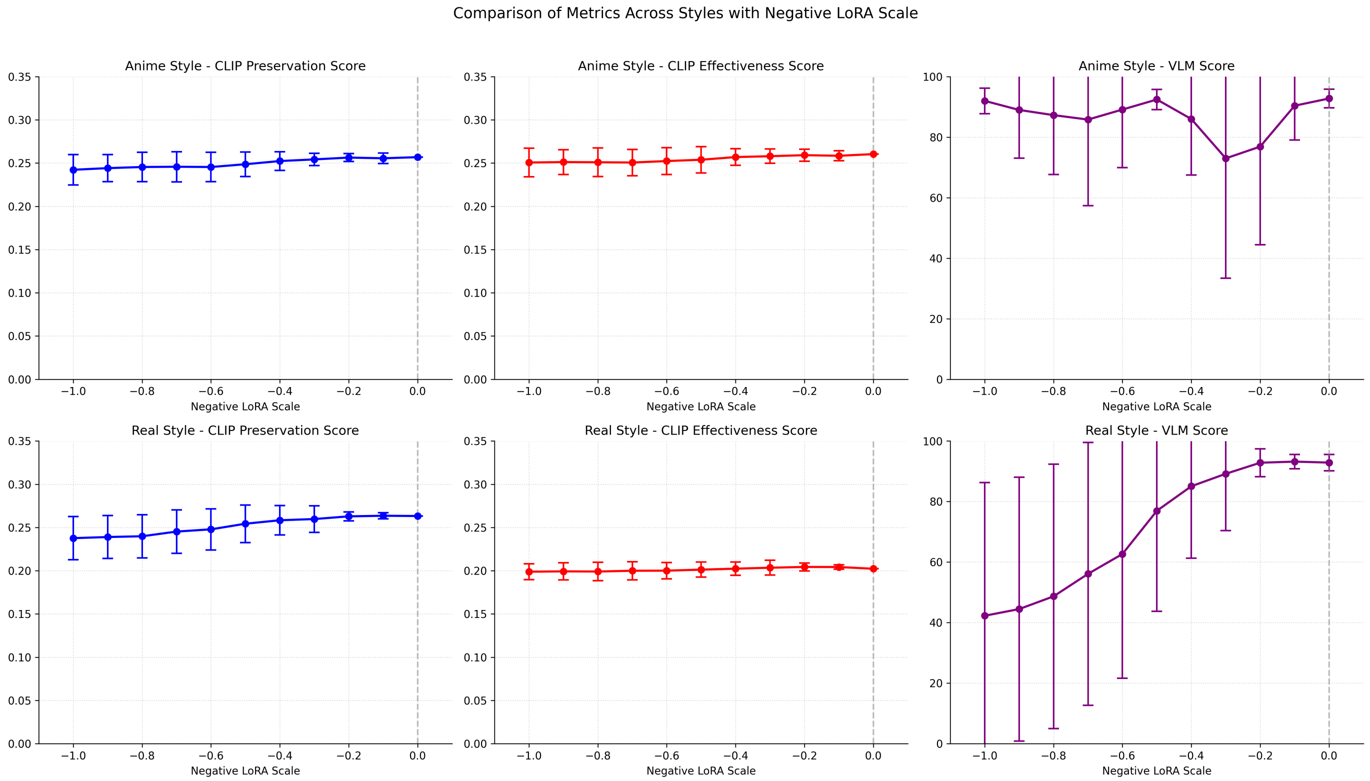

Fig. 1: Comparison of Preservation and Effectiveness Scores Across Negative LoRA Scales for Anime and Realistic Styles: The preservation scores remain stable with minimal deviation, while CLIP and VLM effectiveness scores indicate limited and inconsistent suppression of style, particularly for anime. Realistic style shows a slight trend of style suppression with large variability in VLM scores.

attributes. These evaluations ensure a thorough analysis of both content preservation and style effectiveness.

IV. Experiments

A. *Selection and Labeling of the Dataset*

To unlearn a specific image style, we curated a dataset of LoRA models with diverse styles. Given the constraints of computational resources and efficiency, we utilized Civitai, an open-source platform hosting a wide variety of well-trained models. By scraping data from the Civitai website, we filtered the base models to include only SDXL, yielding a subset of 23,482 LoRA models from a total of 143,000. Upon examining the metadata, we observed a predominance of styles related to human figures, specifically ”anime” and ”realistic.” Thus, we decided to focus our efforts on unlearning these two distinct styles. Initially, binary labeling of models relied on the ”tags” field in the metadata. However, this approach proved noisy, as models tagged ”anime” often lacked anime characteristics in their sample outputs. To address this, we employed CLIP to compute the similarity between the models’ example images and textual descriptions of the target styles. This refined method significantly improved the accuracy and reliability of the dataset labeling, ensuring a cleaner and more precise classification for our unlearning experiments.

B. *Negation of LoRA*

1) Setup: We employ Stable Diffusion XL 1.0 [10] as the base text-to-image diffusion model. To investigate the efficacy of Negation of LoRA, we utilize two sets of pretrained LoRA modules: one trained to amplify “anime” style features and another trained for a “realistic” style. From our filtered dataset, we selected the top 50 anime and realistic LoRA modules based on their CLIP scores, as detailed in Section A. This selection process ensures that only the most effective and representative LoRA modules are incorporated into our experiments, thereby enhancing the reliability and validity of our findings. These LoRA modules are integrated into the model’s cross-attention layers, providing a compact representation of their respective attributes.

To minimize the influence of prompts, we use relatively simple textual prompts (e.g., “a girl, anime” for anime style tests) that do not explicitly encourage any particular disrupting style. We systematically vary α from -1.0 to +1.0 in increments (e.g., 0.2) and generate images at each step. Positive α values reinforce the target attribute, while negative values aim to suppress it, thereby testing whether Negation of LoRA can effectively unlearn the attribute.

2) Result Analysis: Across both the anime and realistic style LoRA models, the preservation scores remain stable as

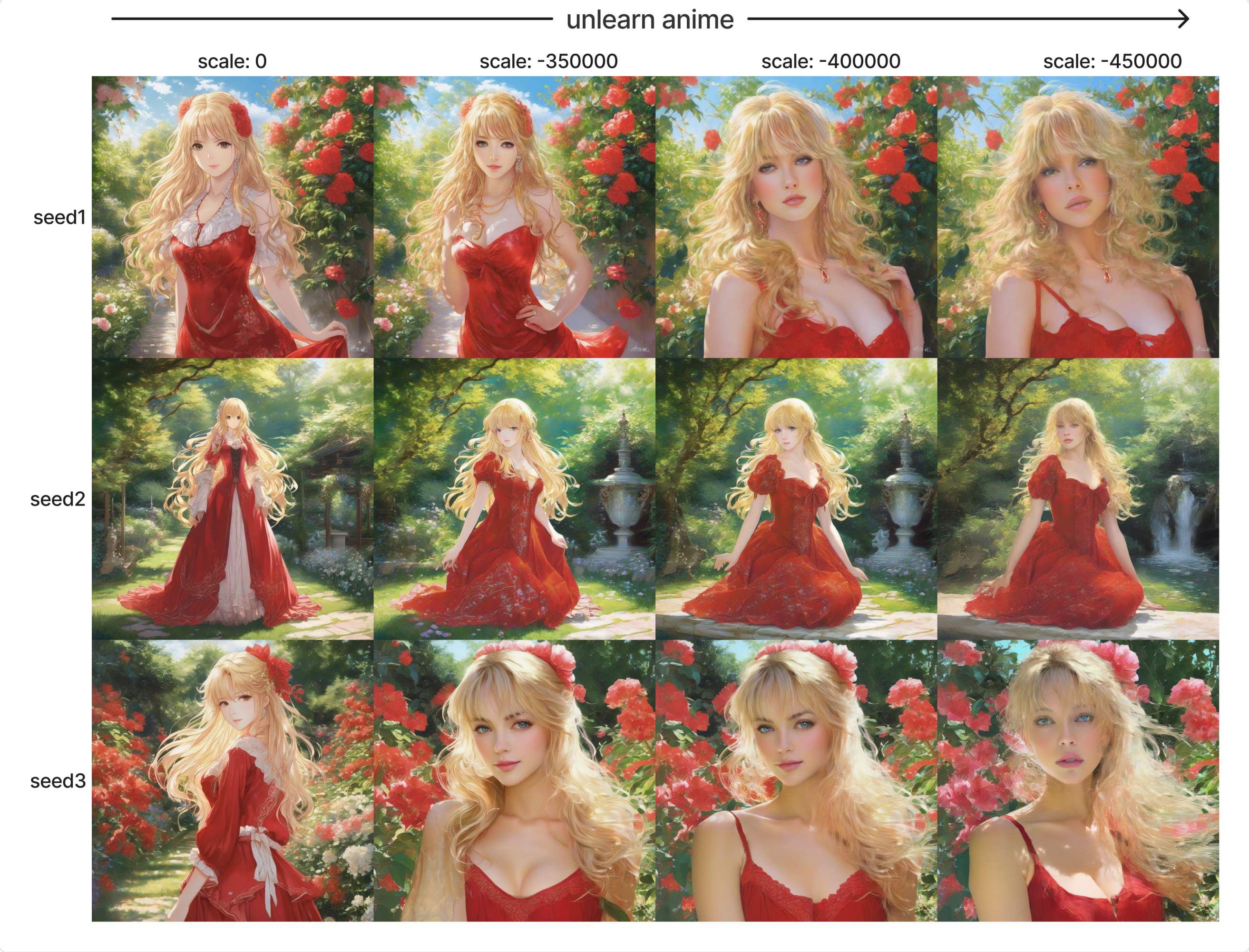

Fig. 2: Progressive unlearning in the W2W space, showing a smooth transition from anime to realistic style as unlearning intensity increases.

the negative LoRA scale increases. As shown in Figure 1, for anime, the preservation scores range narrowly from 0.2423 to 0.2569, while for realistic style, the scores range from 0.2377 to 0.2636. These small deviations indicate that nega LoRA generally maintains the content of the prompt across different scales. Additionally, the preservation scores do not show a clear downward trend as the negative scale increases, further reinforcing the model's stability in retaining content.

The effectiveness scores, measured by CLIP with style-specific prompts, show minimal changes as the negative scale increases. For anime style, the effectiveness score ranges from 0.2604 to 0.2507, with no significant decline observed. Similarly, for realistic style, the effectiveness score ranges narrowly between 0.1989 and 0.2044, again with no clear evidence of stronger suppression at higher negative scales. This suggests that nega LoRA does not effectively reduce stylistic alignment as the scale increases, and its performance in negating style is not substantially based on CLIP effectiveness scores.

The VLM style scores reveal more pronounced but inconsistent changes, particularly for the realistic style. In the realistic style, as the absolute value of the negative scale increases, the VLM scores gradually decrease, indicating some suppression of style. For instance, the VLM score drops from 93.2045 at -0.1 to 42.2727 at -1.0. However, the standard deviations are exceptionally large (e.g., 43.98 at -1.0), which points to highly inconsistent performance. This variability undermines the reliability of style suppression in realistic style LoRA.

For anime style, the VLM scores exhibit more fluctuations and no clear trend of improvement as the scale increases. The scores range from 92.7857 at 0.0 to 73.0 at -0.3, with significant variability at some scales (e.g., 39.59 at -0.3). This inconsistency further suggests that nega LoRA struggles to reliably unlearn stylistic alignment for anime-style prompts.

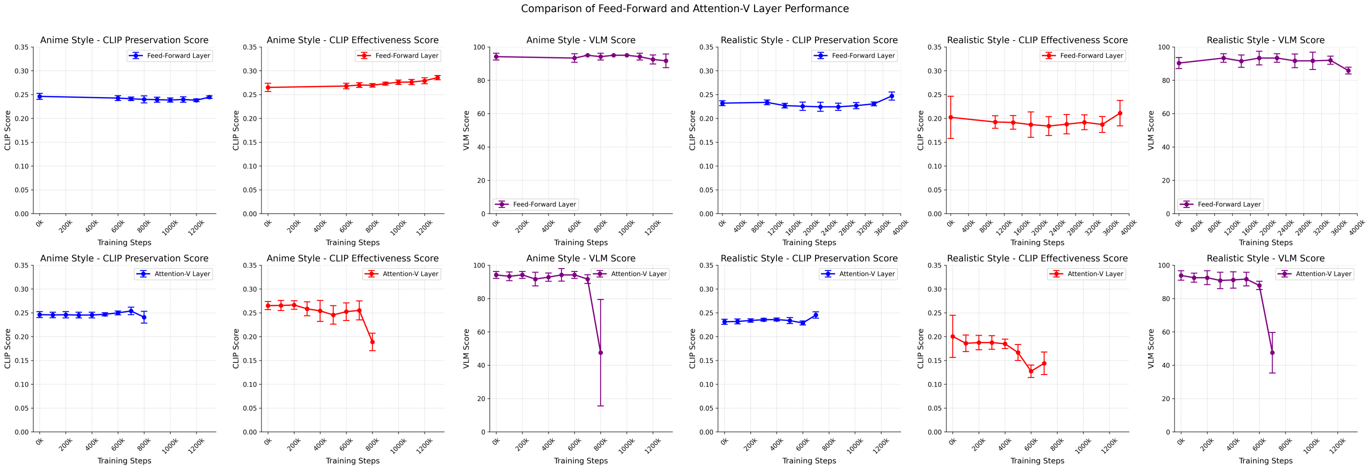

Fig. 3: Comparison of Feedforward and Attention-v Layers across Anime and Realistic styles using CLIP and VLM evaluation metrics. The x-axis represents the unlearning scale, while the y-axis shows the corresponding scores with error bars indicating standard deviation. The first row represents results for Feedforward layers, and the second row represents results for Attention-v layers. From the results, Attention-v layers demonstrate more stable and effective performance in both preserving content and eliminating style attributes compared to Feedforward layers.

C. *Unlearning with W2W Space*

1) Setup: The use of Principal Component Analysis (PCA) for identifying the principal components of the weight space necessitates constraints on the size of the input matrix. Given the variability and noise inherent in publicly generated LoRA models, we implemented two strategies to limit the input matrix size: (i) constraining the rank of selected LoRAs, and (ii) focusing on specific parts of the LoRA model during weight extraction.

In the original W2W paper, the authors utilized self-trained, rank-1 LoRAs. However, LoRAs from the Civitai platform exhibit a diverse range of ranks and structures. To improve computational efficiency for PCA, we filtered the dataset to include only LoRA models with a consistent rank of 16, resulting in a final subset of 857 models.

To further reduce the input weight size, we selectively included certain layers of the LoRA models. To evaluate the significance of different LoRA layers, we conducted experiments where the weight residuals from specific layers were loaded onto base models. Groups of images were generated, and **CLIP scores** were calculated by comparing these images to those generated by the base model using the same seed.

The results (See TABLE 1) revealed that feed-forward (FF) layers had a greater impact on the base models compared to attention value (attn_v) layers. However, FF layers also contain more parameters than attn_v layers. Consequently, in subsequent experiments on weight space learning, we explored both the attn_v and FF layers to balance computational efficiency and effectiveness.

Layer Type	Average CLIP Score	Number of Models
attn_v	0.8851	24
attn	0.8433	24
ff	0.8319	24
ff+attn_v	0.7774	24

TABLE I: Comparison of CLIP scores across different layer types.

2) Result Analysis: **Qualitative Results** The unlearn directions in the W2W space demonstrate compositional properties, allowing for linear interpolation, as illustrated in Figure 2. The first column displays images generated by the unedited SDXL base model, while the subsequent columns show images from progressively edited models with increasing unlearned intensity. Each row shares the same generation seed for consistency. We refer to this weight space as the "anti-real" weight space. This approach reveals a smooth transition where the target style is gradually unlearned. For instance, in the provided example, a girl depicted in an anime style evolves incrementally into a realistic style as the unlearning scale increases. Unlike traditional latent space unlearning methods, which produce isolated edits tied to specific images, the W2W space enables consistent and interpretable style modification at the model level, allowing for the generation of a diverse range of unlearned outputs across multiple instances.

Quantitative Results We evaluated the performance of using specific LoRA layers to construct the weight space, comparing attention-v layers with feedforward layers. As illustrated in Figure 3, Both feed-forward (FF) and attention-v layers perform well in preserving content, with minimal deviations in CLIP preservation scores across training steps. For anime style, FF layers maintain scores between 0.2380 and 0.2462, while attention-v layers show a slightly more stable range of 0.2408 to 0.2539. Similarly, in realistic style, FF layers range from 0.2243 to 0.2470, while attention-v layers achieve consistent scores of 0.2287 to 0.2454. These results

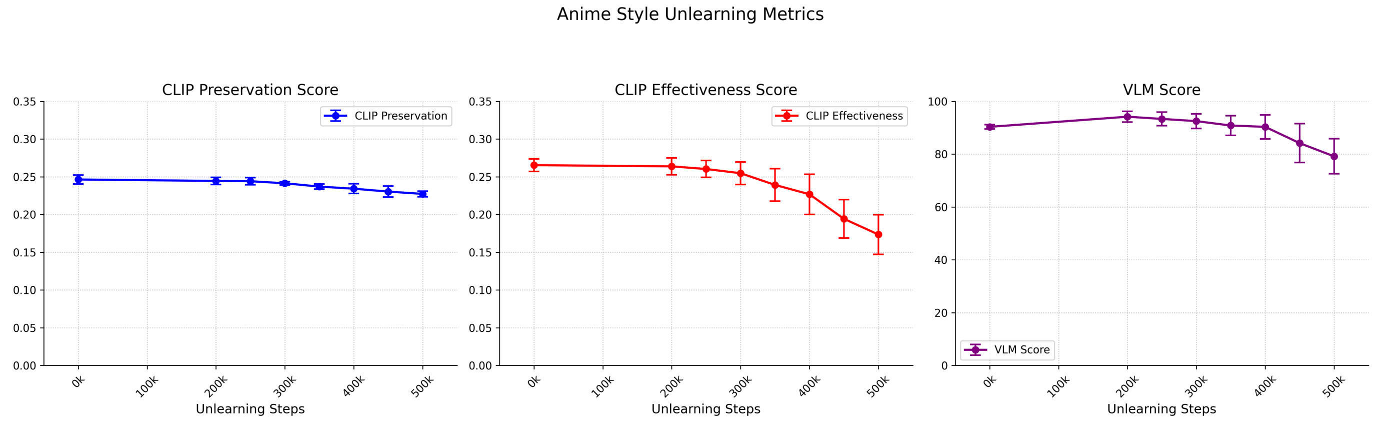

Fig. 4: Evaluation of unlearning anime style using CLIP and VLM metrics on “anti-real” weight space. The x-axis represents the unlearning steps, while the y-axis indicates the corresponding scores with error bars showing the standard deviation. The CLIP Preservation Score (left) reflects the model’s ability to maintain content, where stability indicates better content preservation. The CLIP Effectiveness Score (middle) and VLM Score (right) show the model’s ability to eliminate anime style elements, where lower scores indicate stronger elimination. The results highlight a consistent content preservation and effective unlearning as the unlearning steps increase.

confirm strong preservation by both layers, with attention-v layers showing slightly better. Attention-v layers outperform FF layers significantly in eliminating style. For FF layers, CLIP effectiveness scores show no clear downward trend even at large scales (e.g., 3800k steps), remaining between 0.2650 and 0.2856 for anime and 0.1839 and 0.2110 for realistic style. Similarly, VLM scores remain high across scales, with anime scores between 91.6667 and 95.0000 and realistic scores between 85.8333 and 93.3333, highlighting their limited suppression capabilities. In contrast, attention-v layers show consistent and significant reductions in style alignment at much smaller training scales (around 800k steps). For anime style, CLIP scores decrease from 0.2652 to 0.1888, and VLM scores drop sharply from 94.1667 to 47.5000. A similar trend occurs in realistic style, where CLIP scores drop from 0.2006 to 0.1274 and VLM scores from 93.8333 to 47.5000. This demonstrates the superior efficiency of attention-v layers in eliminating style.

We further evaluated the performance of attention-v layers by constructing an ”anti-real” weight space, a conceptual subspace that captures the transition between ”anime” and ”realistic” LoRA attributes. As shown in Figure 4, The CLIP preservation scores remained stable across unlearning steps, ranging from 0.25 at 0k steps to 0.23 at 500k steps, with minimal deviations.

The CLIP effectiveness scores exhibit a clear downward trend, dropping from 0.26 at 0k steps to 0.17 at 500k steps, highlighting strong and sustained suppression of anime style attributes. Compared to previous experiments, this interpolated weight space achieves a faster and more substantial reduction in CLIP scores (e.g., previous experiments saw scores drop from 0.2652 to 0.1888 over 800k steps, whereas here the scores drop more sharply over just 500k steps). While the VLM scores also decrease steadily from $\sim$ 90 to $\sim$ 70, their decline is less pronounced than in previous experiments. However, the smaller deviations in VLM scores indicate improved reliability, suggesting that the VLM results in this experiment are more consistent and trustworthy. Together, these results showcase the superior ability of the interpolated ”anti-real” weight space to effectively and efficiently eliminate style attributes.

Findings

This paper explores two novel methods for style unlearning in diffusion-based image generation models, leveraging Low-Rank Adaptation (LoRA) modules: Negation of LoRA and Weights2Weights (W2W) unlearning. Negation of LoRA applies negative weights to LoRA modules to suppress specific styles, demonstrating good content fidelity but inconsistent style elimination performance. In particular, it struggles with reliable suppression of anime and realistic styles, as evidenced by significant variability in Vision-Language Model (VLM) evaluation scores.

The Weights2Weights (W2W) framework, on the other hand, extends beyond one-dimensional negation by employing a reduced-dimensional subspace derived via Principal Component Analysis (PCA). This approach consistently achieves robust and interpretable unlearning outcomes, offering smoother transitions in style suppression compared to Negation of LoRA. Furthermore, W2W demonstrates substantial efficiency in eliminating target styles while maintaining stable content fidelity, as measured by quantitative metrics such as CLIP and VLM scores.

Layer-specific insights reveal that attention-v layers are more effective than feedforward (FF) layers in unlearning style attributes. Attention-v layers not only outperform FF layers in terms of style suppression but also maintain higher stability and consistency across training scales. These findings underscore the importance of layer selection when applying unlearning techniques in generative models.

Finally, the paper introduces the concept of an anti-real weight space, which interpolates between anime and realistic LoRA attributes. This weight space enables smooth transitions in unlearning, achieving efficient style suppression with minimal impact on content preservation. The anti-real weight space exemplifies the potential for controlled, interpretable, and scalable style editing at the model level.

These findings highlight the promise of parameter-efficient methods, such as LoRA and W2W, in addressing challenges of style unlearning in diffusion-based image generation. They lay a foundation for future research to further enhance evaluation stability and explore broader applications of these approaches.

V. Limitation

For the evaluation, we chose VLM and CLIP as our primary methods. However, VLM prompting introduces instability, resulting in random score outputs. To mitigate this and enhance accuracy, we designed detailed criteria for VLM scoring, which improved stability but did not completely eliminate inconsistencies. Beyond CLIP and VLM, we explored FID, MSE, and image classifiers to assess negation effectiveness, but these methods proved inadequate for measuring unlearning effects. For FID and MSE, scores increased as the LoRA scale deviated from 0, whether positive or negative, and discrepancies grew between images generated with positive and negative scales. However, these differences appeared to reflect RGB variations rather than stylistic changes, making these metrics unsuitable. We also measured unlearning accuracy (UA) by analyzing changes in the classification accuracy provided by image classifiers. However, for style-related tasks, existing classifiers struggled to differentiate images meaningfully, as negation effects are often subtle, and images generated by the same LoRA were classified into the same style. Moreover, the limited number of images further restricted the reliability of accuracy calculations.

VI. Conclusion

In this study, we addressed the challenges of style unlearning in diffusion-based image generation models, focusing on removing undesired or harmful artistic styles while preserving content fidelity. By leveraging Low-Rank Adaptation (LoRA) modules and introducing two novel methods—Negation of LoRA and Weights2Weights (W2W) unlearning—we provided efficient, scalable alternatives to traditional retraining-based approaches.

Our experiments demonstrated that while Negation of LoRA offers a straightforward approach to style suppression, its performance is limited by one-dimensional traversal in the parameter space, leading to inconsistent results. In contrast, the W2W framework, which utilizes a reduced-dimensional subspace to navigate style-specific weight adjustments, consistently achieved more robust and interpretable unlearning outcomes. Both methods maintained high content preservation scores, highlighting their ability to selectively suppress target styles without degrading the generative model's overall quality.

This work underscores the potential of parameter-efficient frameworks, such as LoRA and W2W, for targeted style unlearning in diffusion models. Our findings pave the way for future research in fine-grained model editing, enabling enhanced control and flexibility in image generation tasks. However, challenges remain in improving evaluation stability and further refining unlearning metrics to capture subtle stylistic changes. Future work should explore broader applications of these methods, including dynamic content regulation and ethical safeguards in generative AI systems.

References

[1] J. Yao, E. Chien, M. Du, X. Niu, T. Wang, Z. Cheng, and X. Yue, "Machine unlearning of pre-trained large language models," 2024. [Online]. Available: https://arxiv.org/abs/2402.15159

[2] J. Chen and D. Yang, "Unlearn what you want to forget: Efficient unlearning for llms," 2023. [Online]. Available: https://arxiv.org/abs/2310.20150

[3] L. Wang, T. Chen, W. Yuan, X. Zeng, K.-F. Wong, and H. Yin, "Kga: A general machine unlearning framework based on knowledge gap alignment," 2023. [Online]. Available: https://arxiv.org/abs/2305.06535

[4] R. Gandikota, H. Orgad, Y. Belinkov, J. Materzyńska, and D. Bau, "Unified concept editing in diffusion models," 2023. [Online]. Available: https://arxiv.org/abs/2308.14761

[5] J. Seo, S.-H. Lee, T.-Y. Lee, S. Moon, and G.-M. Park, "Generative unlearning for any identity," 2024. [Online]. Available: https://arxiv.org/abs/2405.09879

[6] R. Gandikota, J. Materzynska, J. Fiotto-Kaufman, and D. Bau, "Erasing concepts from diffusion models," 2023. [Online]. Available: https://arxiv.org/abs/2303.07345

[7] Y. Zhang, C. Fan, Y. Zhang, Y. Yao, J. Jia, J. Liu, G. Zhang, G. Liu, R. R. Kompella, X. Liu, and S. Liu, "Unlearncanvas: Stylized image dataset for enhanced machine unlearning evaluation in diffusion models," 2024. [Online]. Available: https://arxiv.org/abs/2402.11846

[8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[9] A. Dravid, Y. Gandelsman, K.-C. Wang, R. Abdal, G. Wetzstein, A. A. Efros, and K. Aberman, "Interpreting the weight space of customized diffusion models," 2024. [Online]. Available: https://arxiv.org/abs/2406.09413

[10] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," 2023. [Online]. Available: https://arxiv.org/abs/2307.01952