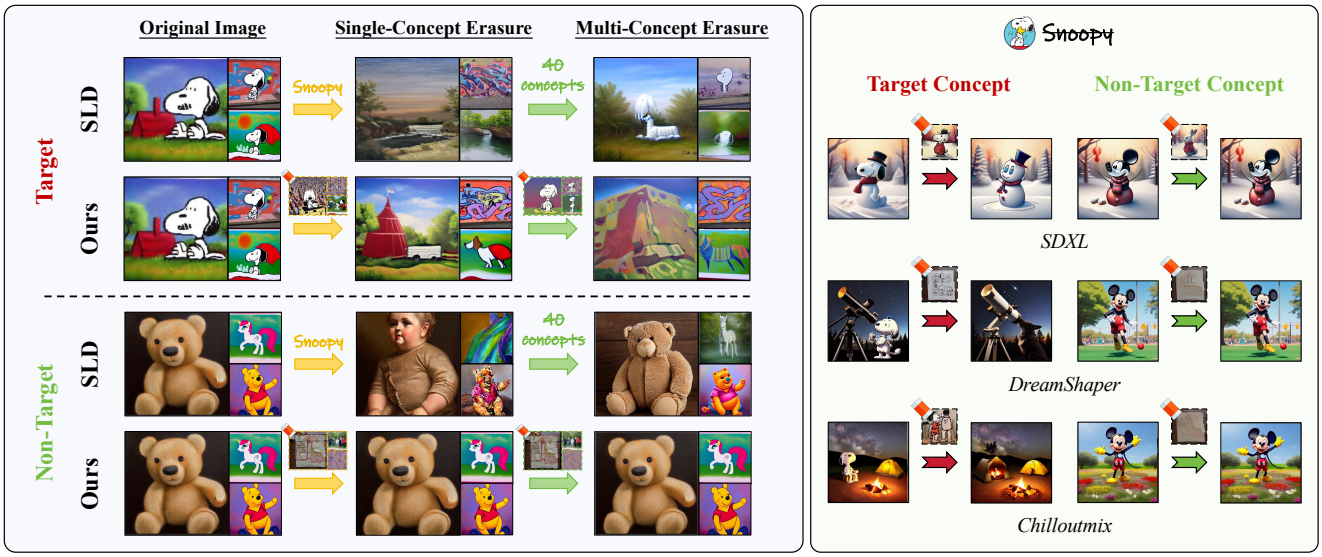
This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Precise, Fast, and Low-cost Concept Erasure in Value Space: Orthogonal Complement Matters

Yuan Wang[1*], Ouxiang Li[1*], Tingting Mu[2], Yanbin Hao[3†], Kuien Liu[4], Xiang Wang[1], Xiangnan He[1†]
[1]University of Science and Technology of China, [2]The University of Manchester,
[3]Hefei University of Technology, [4] Institute of Software Chinese Academy of Sciences
{wy1001, lioox}@mail.ustc.edu.cn, haoyanbin@hotmail.com, xiangnanhe@gmail.com



(a) Erasure Efficacy and Prior Preservation **(b) Transferability**

Figure 1. The proposed Adaptive Value Decomposer (AdaVD) demonstrates a satisfactory balance between erasure efficacy and prior preservation and an effective transferability across T2I diffusion models. (a) Compared to SLD [39], AdaVD enables precise concept erasure without compromising prior knowledge for non-target concepts at both single- and multi-concept erasure. This is facilitated by a precise disentanglement of target semantics (*e.g.*, *"Snoopy"*) and a robust preservation of non-target ones (*e.g.*, *"Teddy"*), with visualization interpretation marked by 🔸. (b) AdaVD can be transferred to various T2I models, e.g., SDXL [31], DreamShaper [6], Chilloutmix [5].

Abstract

*Recent success of text-to-image (T2I) generation and its increasing practical applications, enabled by diffusion models, require urgent consideration of erasing unwanted concepts, e.g., copyrighted, offensive, and unsafe ones, from the pre-trained models in a precise, timely, and low-cost manner. The twofold demand of concept erasure includes not only a precise removal of the target concept (i.e., erasure efficacy) but also a minimal change on non-target content (i.e., prior preservation), during generation. Existing methods face challenges in maintaining an effective balance between erasure efficacy and prior preservation, and they can be computationally costly. To improve, we propose a precise, fast, and low-cost concept erasure method, called **Ada**ptive **V**aule **D**ecomposer (AdaVD), which is training-free. Our method is grounded in a classical linear algebraic operation of computing orthogonal complement, implemented in the value space of each cross-attention layer within the UNet of diffusion models. We design a shift factor to adaptively navigate the erasure strength, enhancing effectively prior preservation without sacrificing erasure efficacy. Extensive comparative experiments with both training-based and training-free state of the arts demonstrate that the proposed AdaVD excels in both single and multiple concept erasure, showing 2 to 10 times of improvement in prior preservation than the second best, meanwhile achieving the best or near best erasure efficacy. AdaVD supports a series of diffusion models and downstream image generation tasks, with code available on: https://github.com/WYuan1001/AdaVD.*

*Equal Contributions.
†Corresponding authors.

1. Introduction

The recent advancements of text-to-image (T2I) diffusion models [11, 17, 18, 30, 36, 50, 52] have enabled users to effortlessly generate high-quality images with simple textual prompts. However, such generations would inevitably introduce copyrighted [10, 22, 42, 44] or offensive [23, 39] concepts, caused by the noisy training data scraped from web [9, 40]. Because it is very costly to re-train large generative models from scratch, it is vital to develop low-cost techniques to precisely erase unwanted semantic concepts in images, *i.e.*, *concept erasure*. This task aims at a precise erasure of visual content w.r.t. target concepts from generated images (*i.e.*, *erasure efficacy*), while a faithful preservation of irrelevant content w.r.t. the prompts comprising non-target concepts (*i.e.*, *prior preservation*), safeguarding a secure T2I generation for diffusion models.

A representative category of concept erasure methods is training-based, which fine-tunes a subset of model parameters by formulating meticulous erasing objective functions [12, 20, 26, 27]. Despite good erasure efficacy, they exhibit considerable practical limitations when being deployed. For instance, they require expensive individual fine-tuning to erase each concept, which thereby limits their real-time usage. As an example, it is unacceptable for online T2I platforms to be costly to erase newly emerging concepts, where copyrighted or offensive concepts could arise unexpectedly, with no means to produce a complete list of concepts to erase in advance. Moreover, these methods suffer from a limited balance between erasure efficacy and prior preservation, due to their reliance on regularization terms to trade off prior preservation.

An alternative category of concept erasure methods is training-free, such as Negative Prompt (NP) [2], Safe Latent Diffusion (SLD) [39], and SuppressEOT [24], which enable real-time erasure. They intervene in the image generation process, exhibiting a range of drawbacks. For instance, NP was initially designed to enhance image quality and can result in compromised erasure efficacy; while SuppressEOT requires the user to specify the location of the target concept within the prompt, thus, it is not suitable for erasure applications that require full automation. Therefore, both NP and SuppressEOT fall short as independent tools for concept erasure. Regarding SLD, it does not perform well at retaining prior knowledge of non-target concepts as illustrated in Fig. 1, failing in *precise concept erasure*. Limited by their drawbacks, current training-free methods are not robust enough to act as an independent concept erasure tool and to be applied in continual erasure of multiple concepts.

In this light, we advance concept erasure techniques for T2I generation, by developing a precise, fast, and low-cost method called **Ada**ptive **V**alue **D**ecomposer (AdaVD). It is a training-free method, capable of precisely erasing the target concepts and satisfactorily preserving non-target priors with low computational overhead. Our core design builds on a classical linear algebraic operation, *i.e.*, projection onto the orthogonal complement of the semantic space of the target concepts. We conduct this projection-based decomposition in the cross-attention value space, disentangling target semantics from the original prompts. To improve the erasure precision and prior preservation, we further refine the decomposition by adaptively allocating token-wise shifts. These shifts are designed to differentiate the strong and specific alignments from the weak and general alignments between prompt tokens and visual content associated with the target concept. Guided by these shifts, the strong and specific alignments are erased for precision while the weak and general alignments are retained for prior preservation.

Fig. 1 (a) demonstrates the erased results for both target and non-target concepts. It shows that our AdaVD can precisely locate and erase the components indicated by the target semantics, meanwhile keeping the non-target priors maximally unaffected. Empirical evaluation shows that AdaVD excels in prior preservation, outperforming the second best by a 2- to 10-fold improvement across various non-target concepts, and meanwhile maintains exceptional erasure efficacy, consistently achieving the best or near-best performance. Our contributions are summarized below:

- A novel and effective erasing operation by exploiting the projection onto the orthogonal complement of the target concept in the cross-attention value space, to disentangle semantics carried by the target concept.
- An adaptive erasing mechanism through a dynamic shift factor, which can effectively minimize the impact on prior knowledge, without compromising the erasure efficacy.
- A precise, fast, and low-cost concept erasure technique AdaVD, which works in a training-free manner and supports a series of T2I diffusion models, capable of precise concept erasure.
- Extensive experiments that demonstrate the superiority of AdaVA against state-of-the-art (SOTA) methods, achieving 2 to 10 times of improvement in prior preservation while maintaining precise erasure efficacy.

2. Related Works

Re-training and Blocking: The most straightforward way to erase a target concept from a pre-trained T2I model is to exclude the training data relevant to this concept and retrain the model from scratch, as in Stable Diffusion (SD) v2.0 [3]. For this, an Not Safe For Work (NSFW) detector [8, 21, 38] can be used to filter unsafe data from LAION-5B [41] prior to the training. However, this approach is time-consuming, requires specialized detectors, and can introduce biases [43]. An alternative solution is to block the prompts of concerns and restrict the outputs of concerns, by using filters [43] and safety checkers [1, 35]. Yet, such safeguards are fragile and easy to bypass [46, 49], especially

when being confronted with crafted malicious prompts [39].

Training-based: A more effective group of concept erasure solutions fine-tune pre-trained generative models, teaching them to "forget" a target concept, e.g., Erased Stable Diffusion (ESD) [12]. However, since the ESD design does not consider prior preservation, both target and non-target concepts are adversely affected. To improve, new training techniques have been developed, *e.g.*, ConAbl [20], SA [14], UCE [13], EraseDiff [48], SPM [27] and MACE [26]. They improve prior preservation through regularization but still fall short in achieving both precise erasure and robust prior preservation. Moreover, for every new target concept to be erased, a separate fine-tuning is required. This is time-consuming, making real-time erasure impractical, especially for highly interactive platforms where unsafe or inappropriate concepts can emerge unexpectedly.

Training-free: With largely reduced computing costs, training-free methods are gaining increasing attention. NP [2] and SLD [39] pioneer training-free concept erasure by adjusting classifier-free guidance [17], leading the generation towards a direction away from the target concepts. However, NP lacks fine-grained control over target concepts and compromises prior preservation. SLD also compromises the prior knowledge during its generation, disrupting the overall generating quality of non-target concepts. SuppressEOT [24], akin to image editing techniques [15], removes target concepts based on user-specified textual positions. Its user-involved design makes it more suitable for editing tasks, but not for system-wide erasing tasks that require full automation. Benefiting from orthogonal decompositions in value spaces, the proposed training-free method AdaVD achieves not only precise concept erasure but also satisfactory prior preservation. It supports multi-concept erasure, is compatible across different versions of stable diffusion, and consistently demonstrates superior performance. The training-free method SAFREE [51], concurrent to ours, also uses orthogonal decomposition but operates differently on text embeddings, supported by masking, projection, Fourier transforms, and a hard control of removal strength. We compare performance with it in the appendix.

3. Method

Concept erasure in T2I generation aims at a successful removal of the visual content indicated by textual concepts (*i.e.*, target concepts), meanwhile a satisfactory preservation of the visual content irrelevant to these concepts (*i.e.*, prior knowledge). It is challenging to achieve simultaneously satisfactory erasure efficacy and prior preservation. To advance this field, we propose a precise, fast, and low-cost concept erasure method, termed AdaVD and illustrated in Fig. 2 (c). The method is training-free, and its core design builds on a classical linear algebraic operation, *i.e.*, orthogonal complement. This simple but elegant geometric operation is able to guide effectively the image generation, enabling a precise removal of the target concepts from the image content meanwhile a satisfactory preservation of the non-target content.

3.1. Preliminary on T2I Diffusion Models

Current T2I models usually include an image compression network [19] and a conditional latent diffusion model [36] that performs sequential denoising with a UNet [37] in the latent space. The UNet takes as inputs the noise latent variable $\mathbf{z}_t$, timestep t, and embedding $\mathbf{C}$ of the textual prompt from a pre-trained CLIP model [33], and predicts the noise $\epsilon_\theta(\mathbf{z}_t, t, \mathbf{C})$. In training-free concept erasure, the noise is additionally conditioned on the target concept with text embedding $\mathbf{C}_t$, predicted by $\epsilon_\theta(\mathbf{z}_t, t, \mathbf{C}, \mathbf{C}_t)$.

Interactions between the image and text modalities are enabled by cross-attention (CA) layers [36, 47] within the UNet, which align the latent representation of the noisy image with the semantic detail of the textual prompt. Each CA layer computes an attention map $\mathbf{A} = \text{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d}}\right)$ with a latent feature dimension d. The queries $\mathbf{Q}$ are computed from the noisy image features, while both keys $\mathbf{K}$ and values $\mathbf{V}$ from the text embeddings $\mathbf{C}$, using different projection matrices. The layer output is a weighted aggregation of $\mathbf{A}$ and $\mathbf{V}$. More details on CA layers are in Appendix A.

It has been recognized that the keys in CA layers mostly act as the "Where" pathway, governing the layout of the attention map and determining the compositional structure of the generated images, while the values as the "What" pathway, controlling the content and visual appearance of images [45]. Because the goal of concept erasure is to modify the visual content of the generated images, we propose to conduct value decompositions into subspaces, which are uniquely constructed by exploiting the target concept and its orthogonal complement in an adaptive fashion. We demonstrate that information offered by the orthogonal complement can be used to generate successfully high-quality images with the target concept precisely erased.

3.2. Token-wise Target Embedding Pre-processing

Given a textual example, which can be either a target concept to erase or an original prompt, its embedding is computed at the token level by a CLIP text encoder. Each tokenized example is padded with `[SOT]` as the prefix and `[EOT]` at the end, with `[EOT]` filling any remaining positions to maintain a fixed token length of l. Each token is characterized by an embedding vector of dimension D_c.

We focus on the embedding of a target concept $\mathbf{C}_t \in \mathbb{R}^{l \times D_c}$, and denote each column of $\mathbf{C}_t^T$ by $\boldsymbol{c}_t^j$ that corresponds to a token vector where $j \in \{1, 2, \ldots, l\}$. To facilitate a precise erasure, we emphasize the key information carried by a target concept by proposing an embedding duplication. Specifically, the embedding of the last sub-

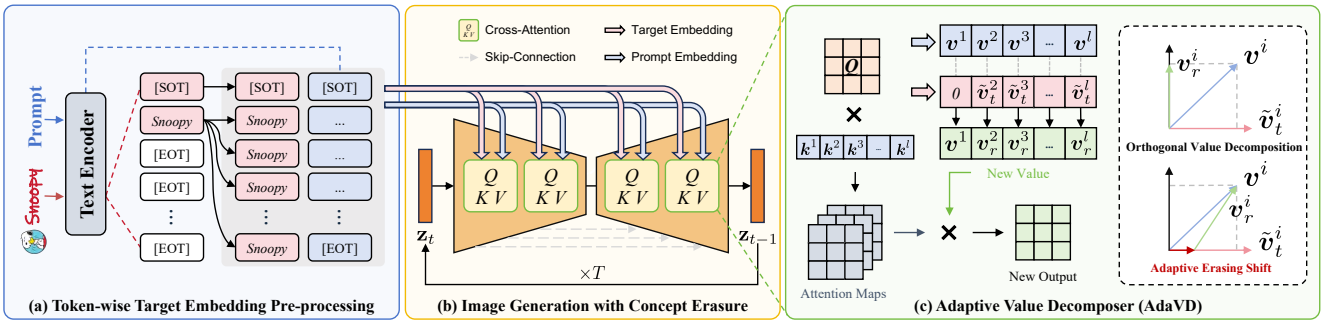


Figure 2. **Overview of our Adaptive Value Decomposer (AdaVD)** in erasing the target concept "*Snoopy*". (a) First, we token-wisely duplicate the last subject token of the target embedding encoded by the text encoder, except for `[SOT]`. (b) Then, the pre-processed target embedding and corresponding prompt embedding are jointly fed into CA layers within the UNet as conditions, to disentangle target semantics from the original image at each timestep. (c) In each CA layer, we perform token-wise orthogonal value decomposition with an adaptive token-wise shift. The new value is subsequently multiplied by the attention map, producing the erased output for this CA layer.

ject token within its prompt content excluding `[SOT]` and `[EOT]`, replaces all the token positions except for `[SOT]`, as illustrated in Fig. 2 (a). For instance, the single-token concept "snoopy" has a modified embedding matrix corresponding to "`[SOT]`, snoopy, snoopy, ..., snoopy", while the multi-token concept "Van Gogh" to "`[SOT]`, gogh, gogh, ..., gogh". Benefiting from the causal attention mechanism used by the CLIP text encoder, the last subject token is able to "see" all the prompt content and contains key information [28], therefore is sufficient for erasure calculation. The modified embedding matrix $\tilde{\mathbf{C}}_t$ is fed into a CA layer to compute the value matrix $\tilde{\mathbf{V}}_t \in \mathbb{R}^{l\times d}$ via a linear projection using the projection matrix $\mathbf{W}_\mathrm{V} \in \mathbb{R}^{D_c\times d}$, as

$$\tilde{\mathbf{V}}_t = \tilde{\mathbf{C}}_t\mathbf{W}_\mathrm{V} = \left[\boldsymbol{c}_t^1, \underbrace{\boldsymbol{c}_t^k, \ldots, \boldsymbol{c}_t^k}_{l-1}\right]^T \mathbf{W}_\mathrm{V}. \tag{1}$$

The token vector $\boldsymbol{c}_t^1$ corresponds to `[SOT]` and $\boldsymbol{c}_t^k$ to the last subject token. The total token number is still l.

3.3. Orthogonal Value Decomposition

Given a conditional latent diffusion model trained for standard T2I generation, our proposed erasing operation works by projecting the original textual prompt onto the orthogonal complement of the subspace spanned by the target concepts to erase, and it is implemented in the value space learned at each CA layer of the UNet. It supports both single-concept and multi-concept erasure.

We do not apply any erasing operation to `[SOT]`, as it primarily serves as a prefix and does not carry useful information to distinguish the semantic content. We start from the modified value matrix $\tilde{\mathbf{V}}_t$ in Eq. (1), and use $\tilde{\boldsymbol{v}}_t^j$ to denote the j-th column of $\tilde{\mathbf{V}}_t^T$. The exclusion of `[SOT]` in the erasure calculation is equivalent to replacing the value vector in the first row of $\tilde{\mathbf{V}}_t$ by a zero vector, resulting in $\mathbf{V}_t = \left[\mathbf{0}, \tilde{\boldsymbol{v}}_t^2, \tilde{\boldsymbol{v}}_t^3, \ldots, \tilde{\boldsymbol{v}}_t^l\right]^T$ to use in the erasing operation.

3.3.1. Single-concept Erasure

Our erasing operation works with $\mathbf{V}_t^T$, where each column of $\mathbf{V}_t^T$ corresponds to the erased value vector for token position j and is denoted as $\boldsymbol{v}_t^j$, and with the value matrix $\mathbf{V} \in \mathbb{R}^{l\times d}$ computed from the original prompt, where each column of $\mathbf{V}^T$ is referred to as the original value vector for the token position j, denoted by $\boldsymbol{v}^j$. To remove the effect of the target concept from the original prompt, for each token position, we project the original value vector $\boldsymbol{v}^j$ onto the orthogonal complement of the span of the erased value vector $\boldsymbol{v}_t^j$, and denote this orthogonal complement by $\mathrm{span}^\perp\left(\boldsymbol{v}_t^j\right)$. This results in the following modified value vector:

$$\begin{aligned}\boldsymbol{v}_r^j = \mathbf{P}_{\mathrm{span}^\perp\left(\boldsymbol{v}_t^j\right)}\boldsymbol{v}^j &= \left(\mathbf{I}_d - \mathbf{P}_{\mathrm{span}\left(\boldsymbol{v}_t^j\right)}\right)\boldsymbol{v}^j \\ &= \boldsymbol{v}^j - \frac{\boldsymbol{v}_t^j\boldsymbol{v}_t^{jT}}{\boldsymbol{v}_t^{jT}\boldsymbol{v}_t^j}\boldsymbol{v}^j = \boldsymbol{v}^j - \frac{\boldsymbol{v}_t^{jT}\boldsymbol{v}^j}{\boldsymbol{v}_t^{jT}\boldsymbol{v}_t^j}\boldsymbol{v}_t^j,\end{aligned} \tag{2}$$

where $\mathbf{P}_\mathbb{X}\boldsymbol{x}$ denotes the orthogonal projection of a vector $\boldsymbol{x}$ onto the space $\mathbb{X}$, and $\mathbf{I}_d$ is an identity matrix of size d. The modified value vector $\boldsymbol{v}_r^j$ is used, instead of $\boldsymbol{v}^j$, to calculate the output of the CA layer, as illustrated in Fig. 2 (c). Since $\boldsymbol{v}_t^1 = \mathbf{0}$, it has $\boldsymbol{v}_r^1 = \boldsymbol{v}^1$, meaning that no erasure is performed for `[SOT]`.

3.3.2. Multi-concept Erasure

We generalize the above operation to erase a set of n target concepts, and their corresponding modified value matrices are denoted by $\left\{\mathbf{V}_t^h \in \mathbb{R}^{l\times d}\right\}_{h=1}^n$. We use $\boldsymbol{v}_t^{h,j}$ to denote the j-th column of $\left(\mathbf{V}_t^h\right)^T$, referred to as the erased value vector for the j-th token position of the h-th target concept. Our erasing operation can naturally be extended to projecting the original prompt to the orthogonal complement of the span of the n erased value vectors $\left\{\boldsymbol{v}_t^{h,j}\right\}_{h=1}^n$, denoted by $\mathrm{span}^\perp\left(\left\{\boldsymbol{v}_t^{h,j}\right\}_{h=1}^n\right)$. To calculate the projection, we first conduct the Gram-Schmidt orthogonalization to obtain a set of n orthonormal basis vectors $\left\{\boldsymbol{o}_t^{h,j}\right\}_{h=1}^n$

for the span of $\left\{\boldsymbol{v}_t^{h,j}\right\}_{h=1}^{n}$. Here, we assume the value vectors in $\left\{\boldsymbol{v}_t^{h,j}\right\}_{h=1}^{n}$ are linearly independent and they form a basis. Such an assumption is reasonable because the multiple concepts to erase should be semantically different in practice, otherwise, a single-concept erasure would be sufficient. The desired projection is then computed by

$$\begin{aligned}\boldsymbol{v}_r^j &= \mathbf{P}_{\text{span}^\perp\left(\left\{\boldsymbol{v}_t^{h,j}\right\}_{h=1}^{n}\right)}\boldsymbol{v}^j = \mathbf{P}_{\text{span}^\perp\left(\left\{\boldsymbol{o}_t^{h,j}\right\}_{h=1}^{n}\right)}\boldsymbol{v}^j \\ &= \left(\mathbf{I}_d - \mathbf{P}_{\text{span}\left(\left\{\boldsymbol{o}_t^{h,j}\right\}_{h=1}^{n}\right)}\right)\boldsymbol{v}^j \\ &= \boldsymbol{v}^j - \sum_{h=1}^{n}\left(\boldsymbol{o}_t^{h,j}\right)^T\boldsymbol{v}^j\boldsymbol{o}_t^{h,j}.\end{aligned} \tag{3}$$

As an addition, we provide in Appendix B an alternative way to compute $\mathbf{P}_{\text{span}^\perp\left(\left\{\boldsymbol{v}_t^{h,j}\right\}_{h=1}^{n}\right)}\boldsymbol{v}^j$ that does not require Gram-Schmidt orthogonalization but matrix inverse.

3.4. Adaptive Erasing Shift

In practice, given a pair of textual prompts and a target concept to erase, their token-wise relevance can vary across different token positions. The different tokens of the prompt carry information with different intensities and focuses and thus can have quite different effects on image generation. As an example, we compare the generated images in each subfigure of Fig. 3, using different versions of the value matrix $\mathbf{V}$ for the same prompt. We separate the values corresponding to the `[EOT]` tokens and the remaining, expressing it as $\mathbf{V} = [\mathbf{V}_{\text{content}}, \mathbf{V}_{\texttt{[EOT]}}]$. Herein, we include `[SOT]` within $\mathbf{V}_{\text{content}}$ for simplicity. We obtain three versions of $\mathbf{V}$ by (1) keeping it as what it is, (2) setting $\mathbf{V}_{\texttt{[EOT]}}$ as zero, and (3) setting $\mathbf{V}_{\text{content}}$ as zero. Fig. 3 shows that the prompt content carries more featured information than those `[EOT]` tokens. This motivates us to further improve the design in Eqs. (2) and (3), by enabling adaptive adjustment of the erasing operation at the token level.

We discover that, although a projection onto the orthogonal complement of the target concept is effective at erasing this concept itself, it can sometimes affect the prior preservation due to an excessive removal of information. Therefore, our adaptive design is focused on improving the prior preservation. When semantics carried by a prompt token are less relevant to a target token, we intend to perform less erasure to protect the prior image content. According to [32], angular information plays a more critical role in conveying semantics than magnitude during image generation. These motivate us to exploit the cosine similarity between the value vectors of a prompt token and a target token and use it to derive a shift factor to adjust the erasing strength along the identified erasing direction, which we refer to as the erasing shift. In general, we let the factor reduce when the cosine similarity becomes smaller.

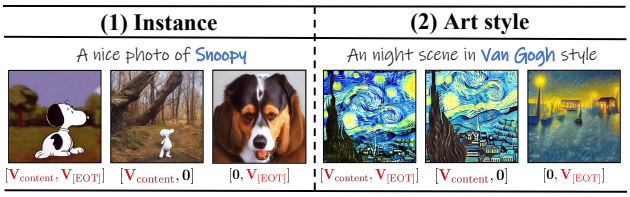


Figure 3. We analyze the contribution of different tokens in text-visual alignment by separately masking the value of content tokens and `[EOT]` tokens, where content tokens carry more featured information than those `[EOT]` tokens.

We denote the shift factor by $\delta(\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, which is a function of two value vectors. Using it, we revise the erasing operation in Eq. (2) for a single concept to

$$\boldsymbol{v}_r^j = \boldsymbol{v}^j - \frac{\delta\left(\boldsymbol{v}_t^j, \boldsymbol{v}^j\right)\boldsymbol{v}_t^{jT}\boldsymbol{v}^j}{\boldsymbol{v}_t^{jT}\boldsymbol{v}_t^j}\boldsymbol{v}_t^j. \tag{4}$$

While for multiple concepts, Eq. (3) is revised to

$$\boldsymbol{v}_r^j = \boldsymbol{v}^j - \sum_{h=1}^{n}\delta\left(\boldsymbol{v}_t^{h,j}, \boldsymbol{v}^j\right)\left(\sum_{k=1}^{n} w_{hk}\left(\boldsymbol{o}_t^{k,j}\right)^T\boldsymbol{v}^j\right)\boldsymbol{v}_t^{h,j}, \tag{5}$$

where w_{hk} is the hk-th element of the projection matrix that transforms the basis $\left\{\boldsymbol{v}_t^{h,j}\right\}_{h=1}^{n}$ to the orthonormal basis $\left\{\boldsymbol{o}_t^{h,j}\right\}_{h=1}^{n}$. For a single target concept, it is straightforward to introduce the shift factor, as in Eq. (4). However, in the case of erasing multiple concepts, the derivation of a shifted operation is less straightforward, as the orthonormal basis does not carry meaningful semantics, thus a transformation back to the value vectors of the target tokens is needed. We explain in Appendix B how to derive Eq. (5) from Eq. (3). The revised erasing operations result in the proposal AdaVD.

Our shift factor design builds on the sigmoid function. Given two input vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we formulate it as

$$\delta(\boldsymbol{x}, \boldsymbol{y}) = \frac{s}{1 + e^{-p(\cos(\boldsymbol{x},\boldsymbol{y}) - \epsilon)}}. \tag{6}$$

The cosine threshold ϵ allows a strong erasure when exceeding the threshold. A negative cosine similarity indicates a very weak relevance between the target and prompt tokens, suggesting no erasure is required. Therefore, we set $0 < \epsilon < 1$ to quantify and filter the relatively weak relevance indicated by a positive but small cosine similarity. The hyper-parameter $s > 0$ controls the factor scale, while $p > 0$ controls the increasing rate of the δ value against the cosine value. In Appendix C.2, we provide hyper-parameter implementation details and a comprehensive analysis.

4. Experiments and Result Analysis

We conduct extensive experiments to evaluate the proposed AdaVD for erasing a diverse range of target concepts, cov-

ering specific instances, art styles, NSFW content, and celebrity. We compare with SOTA training-based methods including ConAbl [20], ESD [12], SPM [27] and MACE [26] and SOTA training-free methods including NP [2] and SLD [39]. We also demonstrate the time efficiency and interpretability of AdaVD, along with its wider usage in other downstream image generation tasks coupled with a series of diffusion models.

4.1. Experimental Setup

Implementation: We employ SD v1.4 [2] to generate images using the DPM-solver sampler [25] over 30 sampling steps with classifier-free guidance [17] of 7.5. All the compared methods are implemented following their default configurations available from their official repository. Further implementation details are provided in Appendix C.1.

Evaluation Data: Adopting the same evaluation protocol from SPM [27], we assess the methods based on 80 instance templates, 30 art style templates, and 25 celebrity templates, and benchmark each method by generating 10 images per template per concept in evaluation. To assess the performance of NSFW erasure, we use the I2P benchmark [39].

Performance Metrics: We follow the widely used evaluation metrics for concept erasure, including the CLIP score (CS) [33] to assess erasure efficacy and the Fréchet inception distance (FID) [16] to assess prior preservation. CS calculates the cosine similarity between a textual prompt and the generated image [33]. We examine two CS values before and after erasing and compare the CS value after the erasure. When the prompts contain the target concepts, a more reduced CS indicates a more effective erasure of the target concepts. FID measures the distance between images generated before and after erasing [16]. A lower FID indicates a better alignment between the two images. Thus, for non-target concepts, lower FID values indicate better prior preservation. Overall, a precise concept erasure should have a low CS for prompts containing the target concepts and a low FID for prompts composed of non-target concepts.

Summary: Results on instance and art style concept erasure are reported in Sections 4.2 and 4.3, while results on celebrity and NSFW erasure are reported in Appendices D.3 and D.2. Additional analyses, including extended quantitative results, comparisons with SuppressEOT [24], performance on different versions of SD, and further discussions, are presented in the Appendix.

4.2. On Instance Concept Erasure

We first conduct the experiment with erasing a single concept *"Snoopy"*. Six types of prompts were tested, of which one contains *"Snoopy"* and the other five contain only non-target concepts. The results are compared in the top block of Table 1. It can be seen that our proposed AdaVD achieves the lowest CS and FID scores in all cases. Particularly, its FID is 33% lower than that of the second-best method. It

Concept	Snoopy	Mickey	Spongebob	Pikachu	Dog	Legislator
	CS	CS	CS	CS	CS	CS
SD v1.4	28.51	26.57	27.43	-	-	-
*Erase **Snoopy***						
	CS ↓	FID ↓	FID ↓	FID ↓	FID ↓	FID ↓
ConAbl	25.38	38.44	41.59	29.68	27.76	27.36
MACE	<u>20.78</u>	118.01	111.90	81.99	43.27	65.97
SPM	23.89	<u>33.06</u>	<u>34.70</u>	<u>23.89</u>	<u>19.61</u>	<u>18.26</u>
NP	23.66	59.58	78.74	52.37	67.51	55.22
SLD	27.84	48.12	55.36	38.74	41.95	49.08
Ours	**20.28**	**5.72**	**8.56**	**5.79**	**2.32**	**6.07**
*Erase **Snoopy** and **Mickey***						
	CS ↓	CS ↓	FID ↓	FID ↓	FID ↓	FID ↓
ConAbl	24.26	24.08	46.32	39.63	30.57	27.49
MACE	<u>20.74</u>	<u>20.71</u>	51.49	110.67	52.07	77.13
SPM	23.16	22.81	<u>41.58</u>	<u>31.77</u>	<u>21.96</u>	<u>23.69</u>
NP	23.59	24.85	81.41	50.10	65.93	58.88
SLD	27.76	26.74	54.59	39.24	41.62	50.13
Ours	**20.29**	**19.93**	**9.34**	**5.84**	**2.41**	**6.43**
*Erase **Snoopy** and **Mickey** and **Spongebob***						
	CS ↓	CS ↓	CS ↓	FID ↓	FID ↓	FID ↓
ConAbl	23.94	23.64	25.04	51.20	31.59	30.03
MACE	<u>20.48</u>	<u>20.50</u>	21.59	99.68	47.46	70.38
SPM	22.81	22.35	<u>20.82</u>	39.83	<u>22.68</u>	<u>25.31</u>
NP	24.29	24.76	25.31	64.75	65.10	59.33
SLD	27.84	26.71	27.60	<u>39.41</u>	42.32	49.88
Ours	**19.39**	**19.73**	**20.34**	**6.85**	**2.79**	**7.26**

Table 1. **Quantitative comparison of single- and multi-instance erasure.** The best and second-best results are marked in **bold** and <u>underlined</u>, respectively. Our AdaVD consistently achieves the lowest CS and the lowest FID in all cases, indicating superior prior preservation without compromising erasure efficacy.

improves over SOTA by significantly enhanced prior preservation without compromising the erasure precision, as exemplified in Fig. 4. It can be observed from Fig. 4 that methods such as SPM, NP, and SLD fail to fully erase *"Snoopy"* which can be found from the ear and the shape characteristic of the generated image after erasure. MACE can successfully erase the *"Snoopy"* concept. But all the competing methods suffer from degraded image quality in non-target concept generation, *e.g.*, *"Spongebob"*.

We then compare the performance for multi-concept erasure, experimenting with two cases, of which one erases two concepts of *"Snoopy"* and *"Mickey"* together and the other erases three concepts of *"Snoopy"*, *"Mickey"* and *"Spongebob"* together. Results are reported in the bottom two blocks of Table 1 and visualized in the second and third sections of Fig. 4. Similarly, our AdaVD achieves the lowest CS and FID scores across all cases. This verifies that AdaVD is capable of handling more complex multi-concept erasure and simultaneously fighting against catastrophic forgetting. Conversely, other methods struggle to consistently perform well for multi-concept erasure. For example, when erasing *"Snoopy"* and *"Mickey"* together, the image quality for non-target concepts like *"Spongebob"* and *"Legislator"* shows apparent degradation in Fig. 4. SLD, in particular, fails to maintain its erasure ability even

Figure 4. **Qualitative comparison of single- and multi-instance erasure.** Both training-based and training-free methods show limitations in prior preservation. In contrast, our AdaVD demonstrates considerable performance in maintaining prior knowledge without compromising erasure efficacy across both single- and multi-concept erasure tasks.

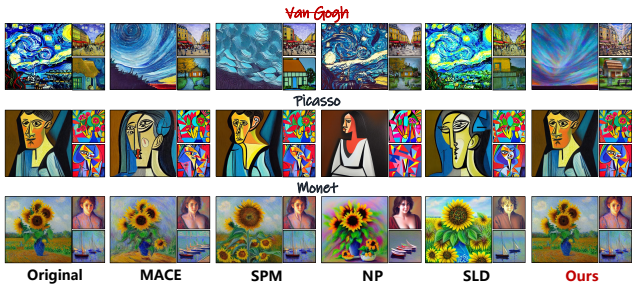


Figure 5. **Qualitative comparison of art style erasure.** Our AdaVD can effectively remove the target concept *"Van Gogh"* while preserving non-target styles like *"Picasso"* and *"Monet"*.

when dealing with 2-concept erasure. In Appendix G, we report more experimental results for erasing up to 40 concepts, where our AdaVD can extend to erase dozens of concepts in practice, maintaining consistent erasure efficacy and prior preservation, as shown in Fig. 1.

4.3. On Art Style Erasure

We experiment with erasing specific art style, including *"Van Gogh"*, *"Picasso"* and *"Monet"*. Results are reported in Table 2, and visual comparisons are provided in Fig. 5. Our AdaVD exhibits superior prior preservation, and achieves the lowest or closed-to-lowest CS and FID scores, demonstrating strong prior preservation without sacrificing erasure efficacy. Other methods show different drawbacks as observed from the results. For instance, although NP achieves notably better precision in art style removal as compared to instance removal, Fig. 5 shows that it still struggles to fully erase the *"Van Gogh"* style. Also, SLD fails to erase the *"Van Gogh"* style. These two methods also degrade generation quality for non-target styles, such as *"Picasso"* and *"Monet"*, showing harmful effects on non-target concepts directly, as evidenced in both Fig. 5 and

Concept	Van Gogh	Picasso	Monet	Andy Warhol	Caravaggio
	CS	CS	CS	CS	CS
SD v1.4	29.21	29.06	29.02	-	-
*Erase **Van Gogh***					
	CS ↓	FID ↓	FID ↓	FID ↓	FID ↓
ConAbl	28.80	71.71	138.72	70.30	73.10
MACE	27.74	65.77	69.79	83.37	75.41
SPM	**24.78**	62.25	32.27	58.30	61.50
NP	24.90	141.56	124.52	127.85	136.32
SLD	27.48	103.96	109.11	103.89	119.32
Ours	24.87	**6.82**	**2.66**	**8.36**	**6.84**
*Erase **Picasso***					
	FID ↓	CS ↓	FID ↓	FID ↓	FID ↓
ConAbl	58.62	27.72	140.34	73.35	67.44
MACE	60.46	27.11	49.92	76.10	72.85
SPM	38.79	26.69	7.76	52.00	51.40
NP	111.35	**26.14**	91.11	116.24	121.82
SLD	98.21	27.03	93.01	97.00	110.05
Ours	**5.49**	26.99	**2.33**	**9.38**	**7.05**
*Erase **Monet***					
	FID ↓	FID ↓	CS ↓	FID ↓	FID ↓
ConAbl	141.52	132.10	24.53	208.38	186.26
MACE	76.90	69.35	26.89	88.35	81.72
SPM	41.03	29.71	27.00	31.90	25.99
NP	137.21	126.75	**24.47**	127.22	135.83
SLD	94.48	92.88	25.73	100.90	114.87
Ours	**6.94**	**6.50**	26.30	**8.46**	**7.19**

Table 2. **Quantitative comparison of art style erasure.** AdaVD achieves a superior balance between erasure efficacy and prior preservation, especially excelling in prior preservation.

Table 2. Both MACE and SPM are effective in erasing the target concept, however, their prior preservation is somehow less satisfactory, which is particularly noticeable in the generated images in *"Monet"* style. Differently, AdaVD can effectively and consistently remove the target concept and meanwhile preserve satisfactorily the prior content, as confirmed by Table 2 and Fig. 5.

	Data Preparation	Model Finetune	Image Generation	Total Time
ConAbl	9290	1120	0.9	10419
SPM	0	72850	1.7	72867
MACE	303	232	0.9	544
SLD	0	0	1.4	14
Ours	4	0	1.8	22

Table 3. **Time consumption of 10-concept erasure**. We calculate the time cost (s) to erase 10 concepts and generate 10 images using one NVIDIA A40 GPU. Compared with training-based methods, AdaVD exhibits exceptional efficiency in real-time erasure.

4.4. Further Analysis

Time Consumption: The computational cost of concept erasure primarily arises from three components, including (1) *data preparation time* required by training-based methods for preparing training data and by AdaVD for basis computation; (2) *model fine-tuning time* required by training-based methods; and (3) *image generation time* required by all methods. In Table 3, we compare the total time consumption of different methods, as well as their time spent on each component. The two training-free methods of SLD and AdaVD are significantly faster as no fine-tuning is needed. Our AdaVD costs slightly more time than SLD, *i.e.*, 0.8 extra seconds per image due to its basis computation, and a total of 8 extra seconds for generating 10 images. But this mild increase yields a significant performance gain, succeeding in precise concept erasure.

Interpreting Erased Components by Visualization: Our AdaVD generates images by replacing the original value vector $\boldsymbol{v}^j$ with $\boldsymbol{v}_r^j$ via orthogonal complement operation. To empirically interpret the rationale behind our method, we visualize the erased component, $\boldsymbol{v}^j - \boldsymbol{v}_r^j$. Fig. 6 presents three cases of erasing the target concepts *"Mickey"*, *"Van Gogh"* and *"Bruce Lee"*, where AdaVD successfully erases these target concepts while robustly preserving prior knowledge of non-target concepts. In the first row, as shown in the middle column of each block, the erased components consistently align with the corresponding target semantics when dealing with the target concepts. In the second row of non-target concepts, conversely, the erased components do not contain any informative pattern, indicating that they carry no meaningful semantics, and therefore exert minimal impact on the prior knowledge.

Downstream Applications: We conduct additional experiments to showcase the versatile applications of our AdaVD across various generative tasks, including (1) implicit concept erasure: by removing the implicit concepts of *"rainy"* and *"foggy"*; (2) image editing: by removing the appearance concepts of *"glasses"* and *"mustache"*; and (3) attribute suppression: by removing the coupled color concept of *"red"*. Additionally, we integrate AdaVD with a series of diffusion models, including Chilloutmix [5], DreamShaper [6], RealisticVision [7], and SD v2.1 [4], alongside SD

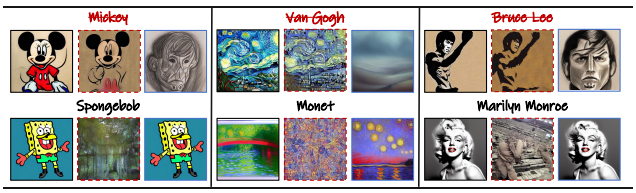


Figure 6. **Visualization of erased components.** In each block, we compare both target (1st row) and non-target concept (2nd row) by visualizing the original image (1st column), erased component (2nd column), and generation by our AdaVD (3rd column).

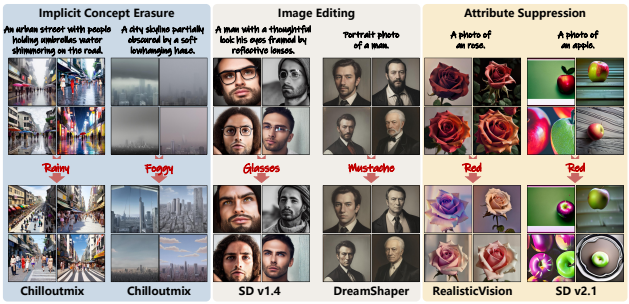


Figure 7. **Downstream applications.** We extend AdaVD to versatile generation tasks, including (1) implicit concept erasure, (2) image editing, and (3) attribute suppression, indicating its significant potential for broad applications.

v1.4. As illustrated in Fig. 7, despite the absence of explicit mention for *"rainy"* and *"foggy"*, AdaVD can still effectively erase these concepts in image semantic space. Meanwhile, AdaVD also precisely removes *"glasses"* and *"mustache"* with minimal changes to other details, highlighting its potential in image editing applications. For attribute suppression, AdaVD successfully eliminates the color attribute *"red"* from objects such as apples and roses, demonstrating its capability to decouple strongly coupled concepts, *e.g.*, "roses are red" embedded in the model's prior knowledge.

5. Conclusion and Future Work

We have presented AdaVD, a precise, fast, and low-cost method for erasing unwanted concepts. The idea of leveraging the classical linear algebraic orthogonal complement operation and an adaptive erasing shift design is novel, and has successfully achieved a precise concept erasure. Extensive experiments have demonstrated both high erasure efficacy and strong prior preservation of AdaVD for both single- and multi-concept erasure. Moreover, AdaVD exhibits excellent interpretability through visualizing its erased components and strong capability in solving downstream tasks. It has been an intriguing discovery that the orthogonal complement of the value vectors of the target concepts is effective at erasing their inherent semantics. Despite the empirical success, we will seek to establish a rigorous theoretical understanding of the accumulative effect of applying this linear algebraic operation layer-wise for concept erasure in the future.

Acknowledgements

This work is supported by the National Science and Technology Major Project (2023ZD0121102) and the National Natural Science Foundation of China (U24B20180).

References

[1] Safety checker nested in stable diffusion. https://huggingface.co/CompVis/stable-diffusion-safety-checker, 2022. 2

[2] Stable diffusion. https://huggingface.co/CompVis/stable-diffusion-v1-4, 2022. 2, 3, 6

[3] Stable diffusion v2.0. https://huggingface.co/stabilityai/stable-diffusion-2, 2022. 2

[4] Stable diffusion v2.1. https://huggingface.co/stabilityai/stable-diffusion-2-1, 2022. 8

[5] Chilloutmix. https://huggingface.co/swl-models/chilloutmix, 2023. 1, 8, 18

[6] Dreamshaper. https://huggingface.co/Lykon/DreamShaper, 2023. 1, 8, 18, 22

[7] Realisticvsion. https://huggingface.co/SG161222/Realistic_Vision_V5.1_noVAE, 2023. 8, 18, 22

[8] Praneeth Bedapudi. Nudenet: Neural nets for nudity detection and censoring, 2022. *URL https://github. com/notAI-tech/NudeNet*. 2

[9] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021. 2

[10] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023. 2

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[12] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436, 2023. 2, 3, 6

[13] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120, 2024. 3

[14] Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[15] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*. 3

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*. 2, 3, 6

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[19] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[20] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702, 2023. 2, 3, 6

[21] Gant Laborde. Deep nn for nsfw detection. 2

[22] Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. *arXiv preprint arXiv:2408.06741*, 2024. 2

[23] Ouxiang Li, Yanbin Hao, Zhicai Wang, Bin Zhu, Shuo Wang, Zaixi Zhang, and Fuli Feng. Model inversion attacks through target-specific conditional diffusion models. *arXiv preprint arXiv:2407.11424*, 2024. 2

[24] Senmao Li, Joost van de Weijer, Fahad Khan, Qibin Hou, Yaxing Wang, et al. Get what you want, not what you don't: Image content suppression for text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*. 2, 3, 6, 19

[25] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022. 6

[26] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2024. 2, 3, 6

[27] Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7559–7568, 2024. 2, 3, 6, 22

[28] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35: 17359–17372, 2022. 4

[29] Carl D Meyer. *Matrix analysis and applied linear algebra*. SIAM, 2023. 11

[30] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 2

[31] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*. 1, 16

[32] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *Advances in Neural Information Processing Systems*, 36:79320–79362, 2023. 5

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 6

[34] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 17

[35] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022. 2

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 3

[38] Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1350–1361, 2022. 2

[39] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023. 1, 2, 3, 6

[40] Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshop Datacentric AI*, number FZJ-2022-00923. Jülich Supercomputing Center, 2021. 2

[41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[42] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2187–2204, 2023. 2

[43] Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of captions. *arXiv preprint arXiv:2006.11807*, 2020. 2

[44] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023. 2

[45] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 3

[46] 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 *The Twelfth International Conference on Learning Representations*. 2

[47] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3

[48] Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. Erasediff: Erasing data influence in diffusion models. *arXiv preprint arXiv:2401.05779*, 2024. 3

[49] Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7737–7746, 2024. 2

[50] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2

[51] Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. Safree: Training-free and adaptive guard for safe text-to-image and video generation. In *The Thirteenth International Conference on Learning Representations*. 3, 19

[52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2