Concept Siever : Towards Controllable Erasure of Concepts from Diffusion Models without Side-effects

Anonymous authors
Paper under double-blind review

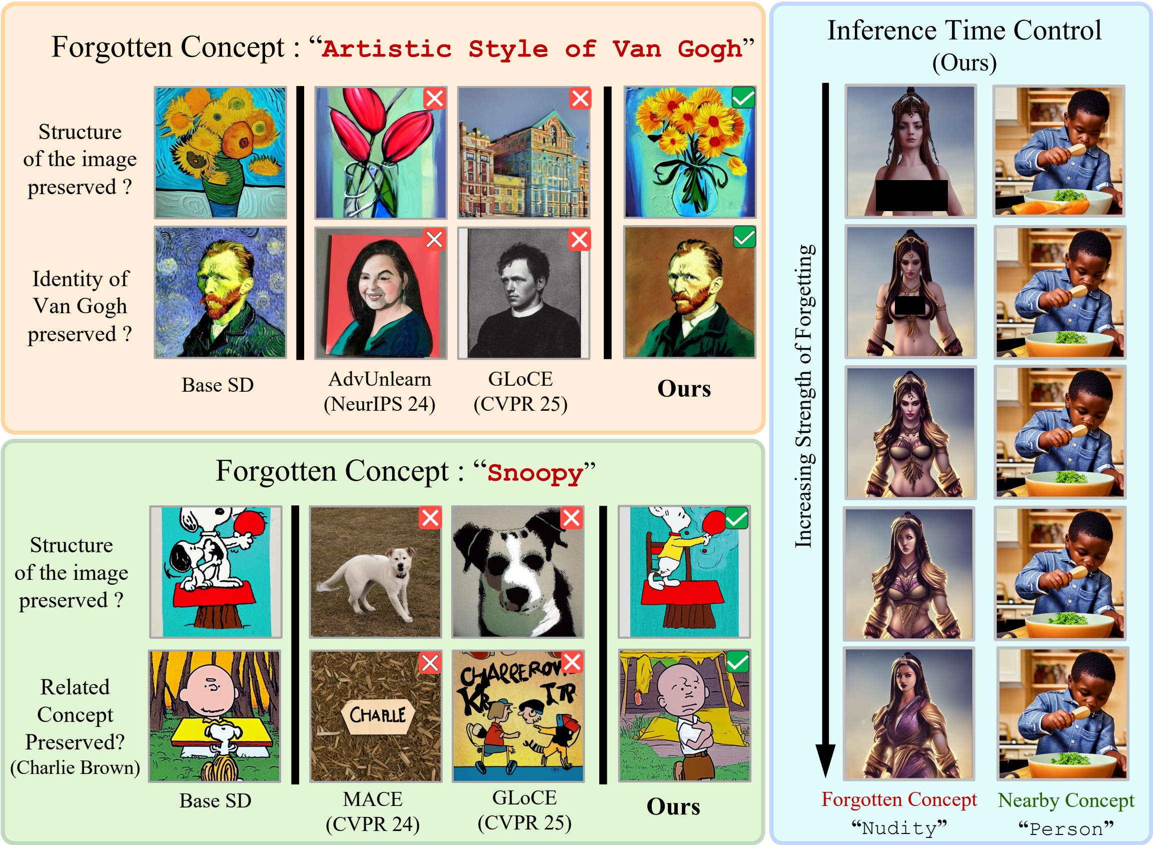
Forgotten Concept : "Artistic Style of Van Gogh"
Structure of the image preserved ?
Identity of Van Gogh preserved ?
Base SD
AdvUnlearn (NeurIPS 24)
GLoCE (CVPR 25)
Ours
Forgotten Concept : "Snoopy"
Structure of the image preserved ?
Related Concept Preserved? (Charlie Brown)
Base SD
MACE (CVPR 24)
GLoCE (CVPR 25)
Ours
Inference Time Control (Ours)
Increasing Strength of Forgetting
Forgotten Concept "Nudity"
Nearby Concept "Person"

Figure 1: We introduce Concept Siever, a concept forgetting framework for diffusion models that provides surgical, concept-level control. By generating a precise vector in weight space, Concept Siever removes a target concept while preserving neighboring ones, overcoming limitations like structure or related concept distortion of state-of-the-art methods like Adversarial Unlearn (Zhang et al., 2025), MACE (Lu et al., 2024) and GLoCE (Lee et al., 2025). Above we show that Concept Siever can forget Van Gogh's artistic style while retaining his identity, and can generate "Charlie Brown" without the unprompted appearance of "Snoopy". A key innovation of our framework is offering fine-grained control over the forgetting strength at inference time, allowing users to dynamically adjust its effect without any retraining.

Abstract

Diffusion models' unprecedented success with image generation can largely be attributed to their large-scale pretraining on massive datasets. Yet, the necessity of forgetting specific concepts for regulatory or copyright compliance poses a critical challenge. Existing approaches in concept forgetting, although reasonably successful in forgetting a given concept, frequently fail to preserve generation quality or demand extensive domain expertise for preservation. To alleviate such issues, we introduce Concept Siever, an end-to-end framework for targeted concept removal within pre-trained text-to-image diffusion models. The foundation of Concept Siever rests on *two key innovations*: First, an automatic technique to create paired dataset of target concept and its negations by utilizing the diffusion model's latent space. A key property of these pairs is that they differ only in the target concept, enabling forgetting with *minimal side effects* and *without requiring domain expertise*. Second, we present Concept Sieve, a localization method for identifying and isolating the model components most responsible to the target concept. By retraining only these localized components on our paired dataset for a target concept, Concept Siever accurately removes the concept with *negligible side-effects, preserving neighboring and unrelated concepts*. Moreover, given the subjective nature of forgetting a concept like nudity, we propose Concept Sieve which provides a *fine-grained control over the forgetting strength at inference time*, catering to diverse deployment needs without any need of finetuning. We report state-of-the-art performance on the I2P benchmark, surpassing previous domain-agnostic methods by over 33% while showing superior structure preservation. We validate our results through extensive quantitative and qualitative evaluation along with a user study.

1 Introduction

Modern large-scale text-to-image (T2I) diffusion models (Rombach et al., 2022; Ramesh et al., 2021; Saharia et al., 2022) are trained on vast amount of internet-scraped data (Schuhmann et al., 2022), raising concerns on data privacy, copyright issues and inappropriate generations like NSFW (not safe for work) images. This has led to development of data protection acts to regulate their usage. For example, established under Article 17 of GDPR (GDPR.eu, 2018), the "Right to be Forgotten" allows individuals to request for deletion of their personal data when it's no longer necessary or if the consent has been withdrawn. Therefore, the *ability to forget* has come up as an important requirement, wherein a model needs to unlearn some particular data it has seen during its training in order to align with the regulatory requirements. Given that these models are trained at large-scale, a *post-hoc intervention* becomes an important necessity to remove specific content when demanded. Towards achieving this, research in *Concept Forgetting* is gaining traction, where one aims to develop mechanisms to identify and remove a particular concept[1] from a model. Doing so effectively can significantly improve these models' reliability, safety, and ethical compliance.

Early concept-forgetting methods (Gandikota et al., 2023; Zhang et al., 2024) have successfully been able to forget a particular concept in a model by designing specialized techniques, but they often suffer from *poor specificity*, where the model's generation capability for neighboring or unrelated concepts get severely impacted. Resolving this inherent trade-off between *forgetting efficacy* and *preserving specificity* is the central challenge of concept forgetting. Recent methods (Zhang et al., 2025; Heng & Soh, 2024; Lu et al., 2024; Gandikota et al., 2024) aim to improve specificity by performing an additional step of preservation using a *preservation set*, which consists of neighboring or unrelated concepts. The nature of this preservation set dictates the degree to which one can improve specificity of the model. For instance, methods like MACE (Lu et al., 2024), UCE (Gandikota et al., 2024) and GLoCE (Lee et al., 2025) typically use a *domain-specific* preservation set — often manually curated — to ensure high specificity. These "concept-aware" sets contain semantically similar examples to the target; for example, to forget one celebrity, the preservation set may include other celebrities who share the common attributes like nationality, ethnicity, or frequent collaborations with the forgotten celebrity. On the other hand, domain-agnostic methods that use generic sets naturally suffer from poor specificity due to the inherent limitations of their preservation set.

Limitations of Domain-specific preservation sets: One of the key limitations of methods using domain-specific sets is their reliance on *domain expertise*, which is often difficult to obtain in practice for users seeking to forget a concept. We demonstrate this issue for GLoCE (Lee et al., 2025), a state-of-the-art domain-specific preservation method, on two forgetting tasks: (a) forgetting NSFW content, and (b) forgetting the celebrity identity of "Brad Pitt". In Fig-2(a), we plot the no. of generated NSFW images by GLoCE when the preservation set shifts from its original, very specific and obscure prompts (e.g., *black modest clothes*, termed

[1] Although there is no universally agreed-upon definition of what is a concept, but generally speaking, a concept can take the form of an object within an image, like a kettle or a human face, to an abstract form, like an artistic style of a painter.

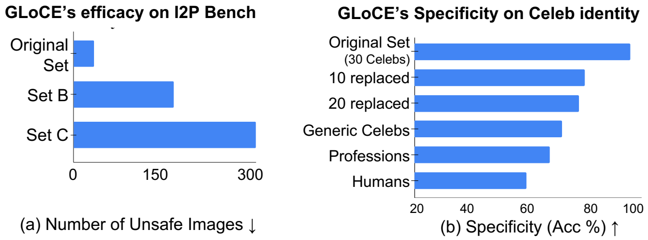

Figure 2: **Sensitivity towards preservation set.** Methods that rely on domain knowledge (here GLoCE is shown) are often highly sensitive to the exact composition of the preservation set, with small progressive changes to it results in sharp drop in performance. See sub-section "Limitations of Domain-specific preservation sets" in Sec-1 for a discussion on this.

Table 1: **Comparison of preservation set curation.** Compared to other methods, Concept Siever satisfies all the desired properties: it *automates* the process of preservation set curation, not requiring any *domain-specific* knowledge to do so, while also being *concept-aware*.

Method	Concept Awareness?	Automated Data Curation?	Domain Agnostic?
UCE (WACV'24)	✓	×	×
MACE (CVPR'24)	✓	×	×
RECE (ECCV'24)	✓	×	×
GLoCE (CVPR'25)	✓	×	×
SA (NeurIPS'23)	×	×	✓
FMN (CVPR-W'24)	×	×	✓
AdvUnlearn (NeurIPS'24)	×	×	✓
Concept Siever (Ours)	✓	✓	✓

"Original" in the plot) to more standard ones (e.g., *a person in modest clothing*, Set C). The observed sharp rise in the generation of NSFW images demonstrates the challenge in designing a perfect preservation set for any given concept. Further, methods that depend on such curated sets often exhibit high sensitivity to the exact set composition. We show an example of this in the case of forgetting celebrity identity of "Brad Pitt" in Fig-2(b), where progressive replacement of the carefully curated celebrity set with increasingly generic celebrities leads to a significant drop in specificity of the model, highlighting the issue of sensitivity. Another key limitation of such domain-specific sets is that they can never be *exhaustive*, i.e., it is practically impossible to curate a set that guarantees the preservation of *all other concepts*. We illustrate a simple example of this in the top-left panel of Fig-1, where forgetting "Artistic Style" of Van Gogh results in erasure of the identity of the painter itself for state-of-the-art preservation-based methods like AdvUnlearn (Zhang et al., 2025) and GLoCE (Lee et al., 2025)[2], demonstrating the inherent limitation and incompleteness of preservation sets.

Unintended side-effects of forgetting algorithms: Specificity is not just influenced by the quality of preservation sets but also by the design of the forgetting algorithm itself. We show multiple examples of this in Fig-1. For instance, in the top-left panel, forgetting "Artistic Style" of Van Gogh using methods like AdvUnlearn and GLoCE results in distortion of the entire structure of the image. A similar effect can also be seen in the bottom-left panel of the same figure, where forgetting "Snoopy" completely alters the structure of the scene. Apart from structure distortion, we also observe that such a forgetting also impacts *related concepts* like "Charlie Brown" (see bottom row of bottom-left panel), even though they are made available in the preservation set. Such side-effects are undesirable, as scenes often contain multiple concepts, and forgetting should ideally target only the specified concept without affecting the rest of the image.

Subjectivity in concept forgetting: Finally, we note that the notion of forgetting a concept such as nudity is inherently subjective and deeply influenced by cultural context. We explain this using the example of Fig-1 (right panel), where we progressively downgrade the level of nudity in the same NSFW image. Now depending on one's socio-cultural background, an individual may perceive the fourth image in the sequence as depicting nudity, while another may not even consider the second image to be inappropriate. Consequently, to provide the flexibility of control during inference time over the degree of forgetting of a concept is not only desirable but also crucial for accommodating the diverse sensitivities of different groups.

Desiderata of concept forgetting: To summarize our discussion till now, a good concept-forgetting method should possess the following desiderata: a) *good specificity*, i.e., to ensure the forgetting is targeted and specific, not impacting any other generation capabilities of the model, b) *good generality*, i.e., the ability to ensure *effective* forgetting of a given concept in all possible contexts, c) *domain-agnostic*, i.e., the ability to forget concepts without the need of specific preservation sets, which not only bring with them their own set of limitations as discussed above, but are also not scalable, and d) *inference-time controllability* to provide a user-level control over the degree of forgetting at inference time to cater to diverse requirements.

[2] We note that this issue occurs in a state-of-the-art method like GLoCE even after preserving a hand picked set of 100 neighboring concepts provided by MACE.

In this work, we present **Concept Siever**, a novel concept-forgetting framework that fulfills all these desiderata for forgetting concepts in Text-to-Image (T2I) diffusion models, achieving effective forgetting in just a few minutes with only the model and a set of prompts. Our approach introduces an *automated* pipeline for generating paired datasets of concept and concept-negated images (Stage I in fig. 3) by leveraging the model's latent space. A key property of these pairs is that *they differ only in the target concept*, enabling forgetting *without side effects* and *without requiring domain expertise*. We utilize this dataset to train a Concept Sieve (Stage-II in fig. 3 and section 3.1) which accurately identifies and isolates the components of the diffusion model responsible for the target concept generation (we quantify this in section 3.2). This targeted isolation ensures *specific* forgetting while preserving performance on unrelated concepts. By controlling the strength of this Concept Sieve using a single hyperparameter λ at inference time (section 3.2), we natively provide fine-grained, user-defined *controllability* over the strength of forgetting. Therefore, by combining all the above components, Concept Siever naturally delivers the specificity, generality, domain-agnosticism, and controllability that define a good concept-forgetting method. We illustrate this in fig. 1, where Concept Siever preserves the overall image structure during forgetting (first row of left panels), minimizes impact on neighboring or related concepts through its enhanced specificity (second row, left panels), and provides fine-grained user-control over NSFW content at inference time (right panel).

We empirically demonstrate the utility and effectiveness of our approach on the popular I2P benchmark of NSFW images (Schramowski et al., 2023), where we achieve a significant improvement of over 33% over the current state-of-the-art forgetting methods like AdvUnlearn (Zhang et al., 2025). We also showcase our forgetting results on the diverse concepts of celebrity identity as well as artistic style. To summarize:

- We introduce Concept Siever, a novel end-to-end framework for concept forgetting in T2I diffusion models, that operates without relying on domain-specific knowledge or external models.
- We propose an novel automated method for generating paired datasets of concepts and their negations by leveraging the diffusion model's latent space corresponding to the concept to be forgotten.
- We also propose Concept Sieve, an accurate localization method to identify and isolate components of the model most relevant to the target concept, enabling precise and effective forgetting.
- Our framework, by construction, facilitates fine-grained *inference-time control* of the strength of forgetting without requiring any additional model finetuning.
- We achieve state-of-the-art results on the I2P benchmark (Schramowski et al., 2023) with a significant improvement of over 33% over prior concept-forgetting methods. Our approach also demonstrates superior structure preservation when compared to all fine-tuning based baselines. We validate this claim through thorough quantitative and qualitative evaluation, along with a user study.

2 Related Work

Concept Forgetting. Recent works in concept forgetting (Zhang et al., 2024; Kumari et al., 2023; Schramowski et al., 2023) addresses the challenge of preservation mainly via two paradigms: those requiring expert-designed, domain-specific prompts to preserve related concepts, and those that operate without domain knowledge. Methods like GLoCE (Lee et al., 2025), MACE (Lu et al., 2024) and UCE (Gandikota et al., 2024) belong to the first category, relying heavily on carefully curated preservation sets tailored to each concept. In contrast, methods like Selective Amnesia fall into the second category, using generative replay with a general set of prompts for retraining after forgetting a concept using the technique of Elastic Weight Consolidation (Kirkpatrick et al., 2017). While Selective Amnesia achieves good forgetting, it suffers from significant concept leakage (Heng & Soh, 2024) due to its use of concept-independent regularization. Different from these methods, which rely on manually curated preservation sets, and thus are inherently limited by the quality and the incompleteness of these sets, we completely eliminate such a need and instead fully automate this generation process by leveraging the latent space of a diffusion model conditioned on the target concept. Another line of work (RECE (Gong et al., 2024), AdvUnlearn (Zhang et al., 2025)) try to enhance the robustness of the forgetting process by iterating through adversarial cycles: first they generate prompt attacks (e.g., via gradient-based text perturbations (Gong et al., 2024)) to expose residual concept traces, then apply methods like UCE (Gandikota et al., 2024) and ESD (Gandikota et al., 2023) to enable successful forgetting of the concept. While this co-training paradigm improves resistance to malicious

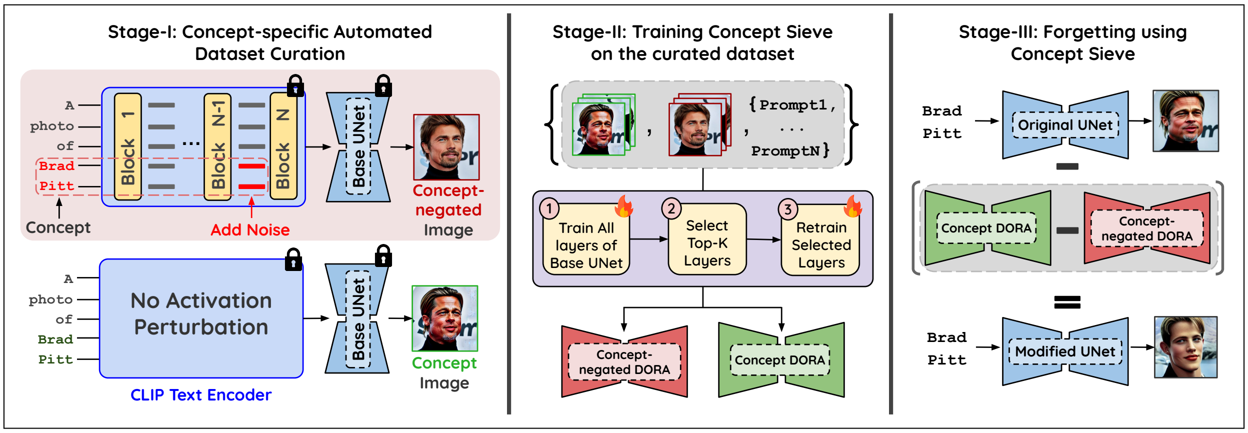

Figure 3: **Concept Siever Framework**: **Stage-I**: We generate paired datasets by perturbing CLIP text embeddings at *concept-specific* token positions, creating concept and concept-negated dataset preserving other attributes. **Stage II**: We train separate DoRA adapters on each dataset and compute their difference as our Concept Sieve (Eq. 4). **Stage-III**: When applied to the original model with scaling factor λ, this vector $\boldsymbol{\tau}$ acts as a filter that selectively removes the target concept while preserving related concepts. It enables precise control over forgetting strength through vector scaling and layer-wise significance scores.

regeneration attempts, it often comes at the cost of reduced specificity. In contrast, our approach effectively retains specificity by leveraging (1) automated concept-aware preservation set, and (2) Concept Sieve, which performs precise identification of the model layers responsible for generating the concept.

Model Editing. Model editing (Yao et al., 2023) has gained significant traction in large language models (LLMs) as a cost-effective alternative (Lu et al., 2024; Orgad et al., 2023) to full model fine-tuning, which often requires substantial computational resources and extensive datasets. Task vectors (Ilharco et al., 2023), a promising model editing approach, deliver strong results with smaller number of training epochs. Their simple arithmetic properties make them composable and suitable for interpolation, enabling flexible adjustments to models (Sanh et al., 2021; Wortsman et al., 2022a;b). But despite their efficacy, task vectors often disrupt model behavior due to their interaction with unrelated portions of the dataset. Additionally, training the entire model can also unintentionally alter its broader functionality. We address these challenges by proposing a sparse and localized technique of identifying target concept within the model components (Concept Sieve, section 3.2), ensuring precise edits without compromising the model's overall integrity.

Efficient Finetuning Methods. Parameter-Efficient Fine-Tuning (PEFT) (Houlsby et al., 2019) refers to techniques designed to reduce the number of trainable parameters, making model training more efficient and cost-effective. A common approach in PEFT methods involves learning parameter changes, $\Delta \boldsymbol{W}$, rather than the full parameter set, $\boldsymbol{W}$ (Hu et al., 2021; Kopiczko et al., 2023; Liu et al., 2024). Among these, adapters based on low-rank decomposition of linear layers have gained popularity. While these methods have made model fine-tuning faster, more accessible, and less computationally demanding, they often become entangled with undesired concepts present in the training data. We address such issues in our proposed method by training a secondary adapter on concept-negated data to capture the collective influence of such undesired concepts, thereby enabling a complete disentanglement of their effects. Subtracting this adapter from the concept adapter yields a well-isolated and disentangled representation direction of the target concept.

3 Proposed Framework: Concept Siever

In this section, we present our proposed framework, shown in fig. 3. At a high-level, our approach isolates the signals that capture the concept to be forgotten from the diffusion model. The only input that is required for this isolation is the name of the concept. The first step is to create a set of paired data $\{\boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{c}}, \boldsymbol{x}_i^{cn}\}$, where $\boldsymbol{x}_i^c$ is an image of the concept to be forgotten and $\boldsymbol{x}_i^{cn}$ is another image with the same characteristics (pose, styling, background etc.) as $\boldsymbol{x}_i^c$, but with the concept negated. We propose a novel automated way for creating such paired data from just the name of the concept, which we explain in section 3.1. Next, we learn a vector in the weight space, which when added to the pre-trained weights, makes the model forget the specific concept of interest. We call this vector as Concept Sieve, explained in section 3.2. It essentially acts as a filter in the weight space that inhibits the concept to be forgotten, while allowing other concepts to pass through. This motivates us to name our methodology as Concept Siever. Concept Sieve, being a vector, allows us to scale it appropriately to have fine-grained inference time control over the strength of forgetting. We note that this flexibility is unique to our novel approach.

3.1 Automated curation of Concept-negated dataset

The core hypothesis of Concept Siever is the ability to isolate the changes in weight space related to the specific concept to be forgotten, while minimizing side-effects on neighboring concepts. Towards this, we first create a dataset of image pairs with and without the concept. A key requirement is that there should be minimal changes in these image pairs except for the concept to be forgotten.

Towards this end, we first curate the concept dataset. Given a concept $\boldsymbol{c}$, say the celebrity Brad Pitt, we obtain prompts with the concepts using LLMs, or using predefined standard templates like "`A photo of Brad Pitt`", "`Brad Pitt on red carpet`". Images generated from Stable Diffusion with these prompts as inputs constitute the concept dataset $\boldsymbol{\mathcal{D}}_c$. To automatically create the equivalent concept-negated dataset $\boldsymbol{\mathcal{D}}_{cn}$, we take the text embeddings corresponding to the concept phrase within the sentence (i.e., embeddings of only the phrase "`Brad Pitt`" in "`A photo of Brad Pitt`") and add Gaussian noise to it:

$$\boldsymbol{t}_i^{cn} = \begin{cases} \boldsymbol{t}_i^c + n \sim \mathcal{N}(0,1), & \text{if } i \in \mathcal{S} \\ \boldsymbol{t}_i^c, & \text{otherwise} \end{cases} \tag{1}$$

where $\boldsymbol{t}_i^c$ are penultimate layer text embeddings from CLIP (Radford et al., 2021) text encoder, $\boldsymbol{t}_i^{cn}$ are corrupted latents, and $\mathcal{S}$ is the subset of indices that are related to the concept to be forgotten. $\mathcal{S} = \{4, 5\}$ for generating the corrupted embedding for the prompt "`A photo of Brad Pitt`". This simple approach, shown schematically in Stage-I of fig. 3, is able to meet our requirements of the two datasets mentioned above. We attribute this localization property to the interpolation-friendly semantic latent space (Bhalla et al., 2024) of CLIP. The noise variance can be varied further to obtain varying degree of separation between the concept and concept-negated dataset. We test different levels noise variance and the layer number of CLIP from which the embedding is extracted in fig. 15 of the supplementary.

3.2 Learning the Concept Sieve

The diffusion model $\phi_{\boldsymbol{\theta}}$ of Stable diffusion (Rombach et al., 2022) operates in the latent space of images. $\phi_{\boldsymbol{\theta}}$ contains a set of encoder and decoder blocks each with self-attention, cross-attention and convolutional layers. Let l denote the number of such layers in the diffusion model. We hypothesize that there exists a Concept Sieve $\boldsymbol{\tau}$ in the weight space, which when removed from $\boldsymbol{\theta}$ would yield a model $\boldsymbol{\theta}^*$ devoid of the specific concept to be forgotten with minimal side-effects to other concepts:

$$\boldsymbol{\theta}^* = \boldsymbol{\theta} - \lambda \boldsymbol{\tau} \tag{2}$$

Inspired by the recent efforts in model editing (Yao et al., 2023) we model $\boldsymbol{\tau}$ as a vector that points from a model trained with forgotten concept, to that of a model trained without the same in the weight manifold, scaled with the hyper-parameter λ. For this, we finetune the base diffusion parameters on the concept and concept-negated dataset. Let $\boldsymbol{w}^i \in \boldsymbol{\theta}$ be the weight in the i^{th} layer of the diffusion model. For finetuning these weights, we adopt a parameter-efficient strategy. Following DoRA (Liu et al., 2024), we choose to decompose $\boldsymbol{w}$ into its magnitude $\boldsymbol{m}$ and direction $\boldsymbol{V}$ components as following (Salimans & Kingma, 2016):

$$\boldsymbol{w} = \boldsymbol{m} \cdot \frac{\boldsymbol{V}}{\|\boldsymbol{V}\|_c} = \|\boldsymbol{w}\|_c \cdot \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_c} \tag{3}$$

Method	Results of NudeNet Detection on I2P (# **of images classified as nude** ↓)									MS-COCO 30K	
	Armpits	**Belly**	**Buttocks**	**Feet**	**Breasts(F)**	**Genitalia(F)**	**Breasts(M)**	**Genitalia(M)**	**Total ↓**	**CLIP ↑**	**FID ↓**
SD v1.4	148	170	29	63	266	18	42	7	743	31.34	14.04
SD v2.1	105	159	17	60	177	9	57	2	586	31.53	14.87
UCE	29	62	7	29	35	5	11	4	182	30.85	14.07
MACE	17	19	2	39	16	2	9	7	111	31.34	13.42
RECE	17	23	0	8	8	0	6	4	66	30.95	14.17
GLoCE*	6	5	2	14	0	0	6	1	34	30.95	13.39
SA	72	77	19	25	83	16	**0**	**0**	292	–	–
FMN	43	117	12	59	155	17	19	2	424	30.39	13.52
AdvUnlearn*	$\underline{8}$	$\underline{5}$	$\underline{1}$	**2**	6	$\underline{1}$	**0**	$\underline{1}$	$\underline{24}$	29.30	15.03
Ours (λ=10)	$\underline{8}$	6	**0**	$\underline{3}$	12	**0**	**0**	2	31	30.34	14.26
Ours (λ=12.5)	**1**	**2**	4	4	**3**	**0**	**0**	2	**16**	29.73	14.51

Table 2: **Forgetting Explicit Content.** Middle panel reports the no. of images flagged as unsafe by NudeNet, with total count reported in the last column. The last panel reports CLIP and FID scores for MSCOCO-30K (Lin et al., 2014) measuring the model's textual fidelity and image quality. Baselines in violet needs domain specific data for training and are not directly comparable. Concept Siever obtain state-of-the-art results in NSFW content forgetting with good preservation quality. In the bottom two rows, we show how scaling down the Concept Sieve with varying λ provides flexible trade-off between preservation quality (last panel) and forgetting efficacy (middle panel). * means the results are adopted from the original paper.

Then, the direction components $\boldsymbol{V}$ are further decomposed into $\boldsymbol{A}$ and $\boldsymbol{B}$ and trained via LoRA (Hu et al., 2021), along with $\boldsymbol{m}$ to obtain the modified weight $\boldsymbol{w}'$ (trainable parameters are underlined),

$$\boldsymbol{w}' = \underline{\boldsymbol{m}} \cdot \frac{\boldsymbol{V} + \underline{\Delta \boldsymbol{V}}}{\|\boldsymbol{V} + \underline{\Delta \boldsymbol{V}}\|_c} = \underline{\boldsymbol{m}} \cdot \frac{\boldsymbol{V} + \underline{\boldsymbol{AB}}}{\|\boldsymbol{V} + \underline{\boldsymbol{AB}}\|_c} \tag{4}$$

When the model is finetuned on datapoints from the concept dataset, we will obtain $\boldsymbol{w}'_c$, and when finetuned on concept-negated dataset, we will obtain $\boldsymbol{w}'_{cn}$. We train both these models independently with the same configuration but with different datasets $\boldsymbol{\mathcal{D}}_c$ and $\boldsymbol{\mathcal{D}}_{cn}$. We define the Concept Sieve $\boldsymbol{\tau}$ as follows:

$$\boldsymbol{\tau} = \{\boldsymbol{w}'^{i}_{c} - \boldsymbol{w}'^{i}_{cn}\}_{i=1}^{l} \tag{5}$$

The objective function $\mathcal{L}$ for this fine-tuning stage is the MSE loss between $\boldsymbol{z}$ and $\boldsymbol{z'}$, where $\boldsymbol{z}$ is the encoded latent representation of the image $\boldsymbol{x}$, and $\boldsymbol{z'}$ is computed as follows: $\boldsymbol{z'} = \phi_{\boldsymbol{\theta}}(\boldsymbol{x}) + \boldsymbol{\tau}^{\cdot}\nabla\phi_{\boldsymbol{\theta}}(\boldsymbol{x})$ [3]. It can also be interpreted as a linear approximation to eq. (2).

Benefits: Concept Sieve $\boldsymbol{\tau}$ offers key advantages to the problem setup of concept forgetting with minimal side-effects: Firstly, it helps to pin-point the specific layer of the diffusion model that has the most affinity towards a task. This score can be computed as follows: $S_{layer} = \|\boldsymbol{m}_c - \boldsymbol{m}_{cn}\|_2 \;/\; d$, where $\boldsymbol{m}_c$ and $\boldsymbol{m}_{cn}$ are the magnitude vectors of the DoRA parameters that we learn in eq. (4) and d is dimension of the layer, while training on concept dataset and concept negated dataset. Empirically we find that if we identify the most important layer(s) using the score S_{layer} and finetune just those layer(s), we can reliably boost the sieving ability of our approach. Further, we can analyze the direction vectors inside $\boldsymbol{V}$ matrix to further prune those columns that maximally effect the erasure of the concept. This provides an extra degree of *inference-time control* without any fine-tuning. We refer to this as *Column Masking.* Its score can be computed as follows:

$$S_{column} = \frac{\boldsymbol{v}_c^i \cdot \boldsymbol{v}_{cn}^i}{|m_c^i - m_{cn}^i|} \tag{6}$$

where $\boldsymbol{v}^i$ is the direction and m^i is the magnitude of i^{th} column vector in $\boldsymbol{V}$ and $\boldsymbol{m}$ respectively.

[3] This equation is inspired from Taylor's first degree approximation, and is applicable in our setup as we want to steer the model towards targeted forgetting gently (Ortiz-Jimenez et al., 2024).

Concept: Celeb Identity		**Forgetting Efficacy**		**Specificity (Other Celebrities)**		
Method	**Domain Agnostic?**	**Top-1 Acc (%) ↓**	**#imgs w/o faces ↓**	**Top-1 Acc (%) ↑**	**#imgs w/o faces ↓**	**LPIPS ↓**
SD v1.4	-	94.74	3	96.61	21	-
UCE	×	0.00	8	95.46	32	0.347
MACE	×	0.83	8	78.09	67	0.543
RECE	×	0.00	23	52.05	83	0.347
GLoCE	×	2.00	3	96.45	21	0.002
SA	✓	20.00	**0**	88.42	**4**	0.563
FMN	✓	18.75	10	64.30	63	0.545
Ours (40% CM)	✓	15.04	4	**92.23**	29	**0.127**
Ours (45% CM)	✓	4.05	3	86.83	32	0.140
Ours (70% CM)	✓	**0.82**	5	75.86	31	0.178

Table 3: **Forgetting Celebrity Identity.** Results on the efficacy of forgetting are reported in the middle panel of the table ("Forgetting Efficacy"), while the last panel reports specificity using GCD accuracy (↑) for 100 unseen celebrities and other metrics. Best results are shown in **bold**, with next best results underlined. Baselines which are not domain-agnostic (×) are not directly comparable. Concept Siever demonstrate state-of-the-art results among the domain-agnostic methods. Last three rows shows our efficacy-specificity control by varying the percentage of column masking (xx% CM) (see section 4.4 and appendix G.4).

4 Experiments and Results

We evaluate Concept Siever on three datasets: **I2P Benchmark** which consists of NSFW content (Schramowski et al., 2023), **Celebrity Identity** (Heng & Soh, 2024) and **Artistic Style** (Gandikota et al., 2023). We evaluate against *seven baselines*, out of which four of them depend on explicit preservation of domain-specific concepts – GLoCE (Lee et al., 2025), UCE (Gandikota et al., 2024), MACE (Lu et al., 2024), RECE (Gong et al., 2024), and the other three which are domain-agnostic – Forget Me Not (FMN) (Zhang et al., 2024), Selective Amnesia (Heng & Soh, 2024) and AdvUnlearn (Zhang et al., 2025). As stated earlier, our method belongs to the latter category, where we not require access to any domain knowledge to curate the preservation set. Different from the preservation set, we curate the *concept dataset* using 40-50 text prompts that are either human-designed or generated by a large language model (LLM), each explicitly referencing the concept to be forgotten. Following previous works (Heng & Soh, 2024; Gandikota et al., 2024; Gong et al., 2024; Zhang et al., 2025), we conduct all our experiments using Stable Diffusion v1.4 (Rombach et al., 2022) as the base diffusion model. We run the DDIM (Song et al., 2020) sampler for 50 time steps.

4.1 Forgetting Explicit Content

We follow MACE's evaluation protocol for benchmarking on I2P dataset for forgetting NSFW content. Detailed evaluation procedure is provided in appendix F.1. The results are presented in table 2, where we can observe that Concept Siever sets a new state-of-the-art performance in removing NSFW content, improving over prior domain-agnostic SOTA methods like AdvUnlearn (Zhang et al., 2025) by a significant margin of over 33%. We further demonstrate fine-grained inference-time control over the NSFW content in Fig-1 (right-panel) by varying the λ parameter (eq. (2)). Such a control can cater to diverse sensitivities of different sub-groups at once, without the need of any model finetuning. Moreover, our approach also maintains comparable or better semantic knowledge and image quality compared to existing domain-agnostic baselines, as demonstrated by the CLIP and FID scores in the same table. For reference, we also include SD v2.1 as a baseline in table 2, as it is the SD version fine-tuned on training data with NSFW content filtered out.

4.2 Forgetting Celebrity Identity

Following Selective Amnesia (Heng & Soh, 2024), we aim to forget the identity of the actor "`Brad Pitt`" from Stable Diffusion, and evaluate the results using the standard GCD classifier (Heng & Soh, 2024) (lower

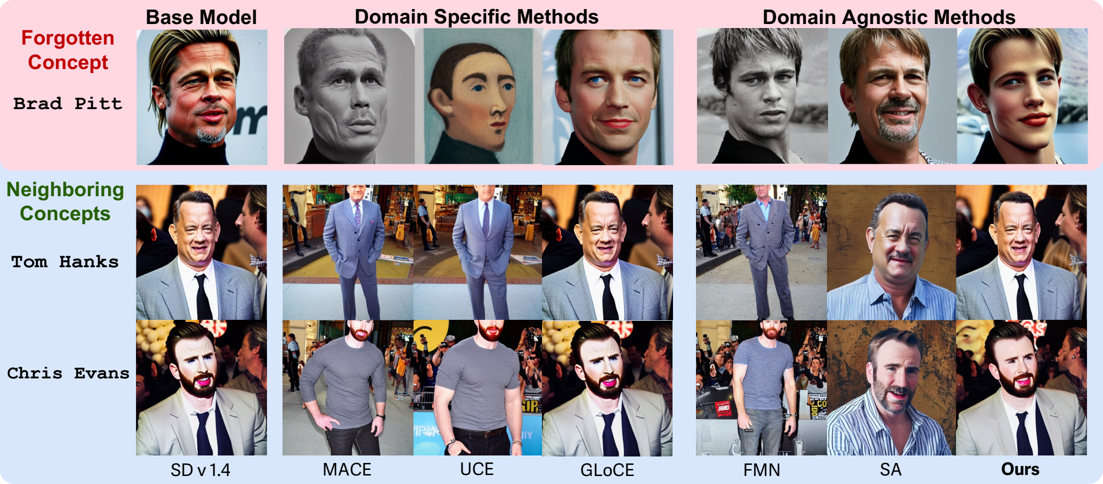

Figure 4: **Forgetting Celebrity Identity.** Post Forgetting "Brad Pitt" (top row), Concept Siever (last column) exhibits significant improvement over Selective Amnesia (SA). Note that methods like MACE and UCE generate images without faces for neighboring concepts (see bottom panel).

value of this metric implies better forgetting). More details on the evaluation procedure and metrics can be found in appendix F.2. The results are shown in table 3. From the table, we report superior Top-1 Accuracy results in the forgetting efficacy (see "Forgetting Efficacy" panel, lower is better here, implying better forgetting) compared to all the existing domain-agnostic baselines. We also report two more results of ours corresponding to the different column masking fractions to demonstrate better specificity-efficacy trade-off.

To assess the specificity, we report Top-1 Accuracy, number of images where the model does not generate any face after forgetting (higher number implies poor specificity), and LPIPS metric for 30 other celebrities. The results of these metrics are also shown in the same table (table 3). As one can observe, Concept Siever causes minimal impact on the model's ability to generate faces (imgs w/o faces ↓), or any other image characteristics (LPIPS ↓). We also present qualitative results in fig. 4, where we observe the same trend — better preservation capabilities of Concept Siever compared to baseline approaches. This is facilitated by precise layer localization of our framework which we discuss in detail in section 5.2. Additionally, the "concept awareness" of our automated preservation set also allows us to effectively preserve neighboring concepts despite being domain-agnostic. Additional qualitative result comparisons are provided in appendix L.1.

4.3 Forgetting Artistic Style

Following ESD (Gandikota et al., 2023), we evaluate the forgetting of the "Artistic Style" of Van Gogh. For methods requiring explicit preservation (GLoCE, UCE, RECE, MACE), we preserve 100 different artistic styles for fairness. Standard evaluation for artistic style uses LPIPS measure between the original and forgotten images (we term this as "Naive LPIPS"); however, *this metric can be misleading, as large structural distortions — undesirable as per our desiderata mentioned earlier — will also contribute to a higher LPIPS score, which is incorrect.* Therefore, we separately quantify structure preservation by observing that structure (or content) in an image is primarily represented in the low-frequency components, and therefore to quantify this, we blur both the original and forgotten SD images using Gaussian noise across varying strengths, identify the blur range where high-level structure remains intact, and compute the average LPIPS between the corresponding blurred images over this range, reflecting the overall degree of structure preservation.

The results are shown in table 4, with the structure preservation result reported under the column "Structure LPIPS". From the table, we note that Concept Siever demonstrates best preservation quality (last panel) along with superior structure preservation among methods that do not explicitly preserve the neighboring concepts. We also present qualitative results in fig. 5, where Concept Siever not only excels at preserving the

Method	Domain Agnostic?	Forgotten Concept: Artistic Style of Van Gogh		Other Concepts
		Naive LPIPS ↑	Structure LPIPS ↓	LPIPS ↓
UCE	×	0.718	0.626	0.271
MACE	×	0.707	0.603	0.437
RECE	×	0.736	0.636	0.398
GLoCE	×	0.722	0.522	0.007
SA	✓	**0.750**	0.818	0.685
FMN	✓	<u>0.727</u>	<u>0.605</u>	<u>0.467</u>
AdvUnlearn	✓	0.745	0.645	0.500
Ours	✓	0.645	**0.442**	**0.228**

Table 4: **Forgetting Artistic style.** We show superior performance of Concept Siever compared to state-of-the-art domain-agnostic methods shown by the columns "Structure Preservation" and "Style LPIPS Difference" for forgetting, and by "Other Concepts" panel for preserving other concepts (specificity). We report specificity by calculating LPIPS on 2500 images related to 100 artists, and efficacy by evaluating on 250 images on the forgotten concept i.e. "Artistic Style" of Van Gogh. Methods with domain-knowledge preserve 100 concepts. Best results are shown in **bold**, with next best results <u>underlined</u>. See Sec-4.3 for more details.

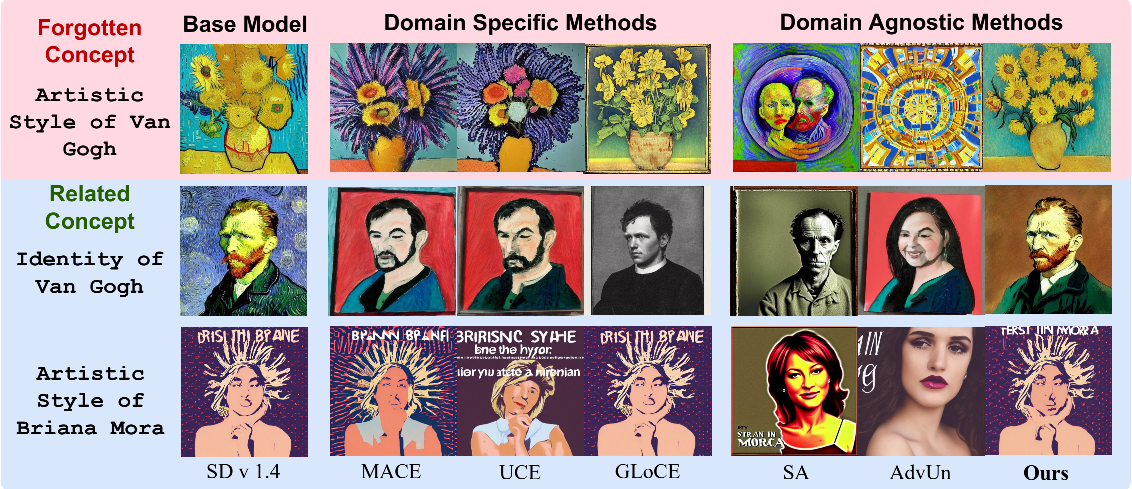

Figure 5: **Forgetting Artistic Style.** After forgetting the artistic style of "Van Gogh" (first row), Among all the baselines, Concept Siever show better preservation of image content in both forgotten and rest of the styles, while maximally preserving the artistic style for the neighboring artist "Briana Mora".

neighboring concepts but also the original structure of the forgotten image, while successfully forgetting the style. To further validate this, we perform a user study with 33 participants on 22 generated samples, asking them to pick the method that gives the best erasure of artistic style *while preserving the image structure and other concepts*. Results show that users prefer our method $\sim 73\%$ of the time compared to the baselines, validating our superior preservation efficacy. This can be attributed to the novel concept-negated dataset as well as the column masking feature of our framework, which precisely identifies the model layers responsible for the target concept, thereby incurring minimal impact on other concepts.

4.4 Inference-time Control over Forgetting strength

As mentioned before, a significant advantage of our framework is the ability to provide fine-grained continuous control over the strength of forgetting at inference time, allowing users to seamlessly navigate the efficacy-specificity trade-off without any retraining overhead. This control is achieved through two complementary mechanisms: (1) scale of the steering vector λ (eq. (2)) and (2) column masking (eq. (6)). Our primary mechanism for control is the learned steering vector τ, which shifts the model's weights to erase a concept.

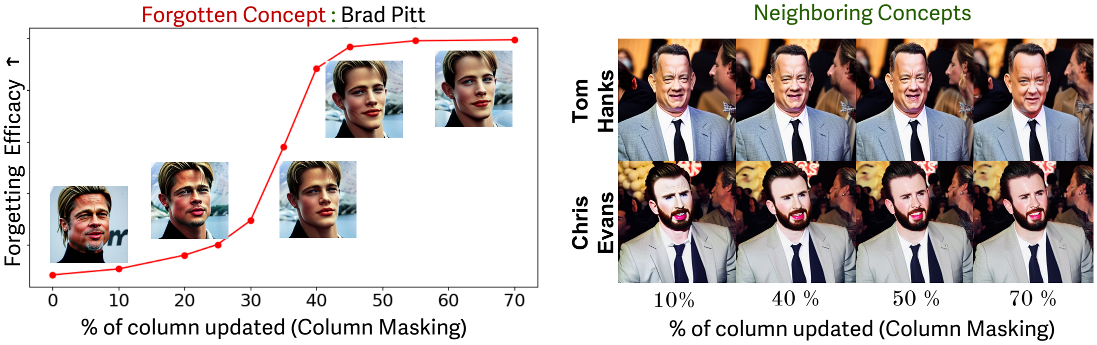

Figure 6: **Inference-time Control using Column Masking.** By slowly increasing the number of columns per layers (left-to-right), we perform successful fine-grained forgetting of Brad Pitt (reflected by efficacy score), with minor impact in neighboring concepts (right plot).

The intensity of this shift can be continuously modulated by λ: By adjusting λ at inference time, a user can smoothly transition from the original model behavior ($\lambda = 0$) to complete concept erasure. We demonstrate this capability qualitatively for NSFW content in fig. 1 (right-panel).

We also leverage column masking (section 3.2) as a complementary second axis of control. This technique allows us to define the *scope* of the intervention by selecting a specific percentage of the most relevant model parameters to update. To demonstrate this control, we conduct an experiment on the concept of celebrity identity "Brad Pitt", where we progressively increase the percentage of updated columns from 10% to 70%. As illustrated in fig. 6, increasing the column mask percentage strengthens *forgetting efficacy*, successfully leading to a corresponding reduction in the classifier's softmax score for the target identity. However, as expected, this increased efficacy produces collateral effects on specificity: At higher masking percentages, we observe a noticeable degradation in the generated identities of neighboring concepts (e.g., "Tom Hanks" and "Chris Evans"), highlighting the direct trade-off that can be managed through this control. Finally, a joint study analyzing the combined effects of the steering vector scale λ, and the column masking percentage for the artistic style domain is provided in the supplementary material (appendix L.2 and appendix I).

5 Method Insights and Analysis

5.1 Targeted Forgetting: Top-K Layer Selection and Column Masking

One of our core hypothesis is that sparse, targeted edits to the model would allow us to preserve its behavior for other concepts. Therefore, to identify these components, we leverage our automated data curation method to guide the editing process (section 3.1). While our data curation method does preserve the image semantics well, some distortions still occur. However, we note that these distortions are unique to each image, whereas the concept-related differences remain consistent across the dataset. Therefore, by training on multiple images, we reduce the leakage significantly. We can further stem the leakage by reducing the no. of selected layers while performing top-K layer selection for finetuning, as well as using column masking (CM) (see Stage-II in fig. 3). To demonstrate this, we perform an ablation on these components and show their results in fig. 7. It is evident from the last two columns of fig. 7 that localizing top layers (top-K) and using column masking (CM) further preserves the image structure and content.

5.2 Concept Localization: Analysis of Cross-Attention Layers

Cross-attention layers play a crucial role in concept representation as they are the conditioning layers for the text prompts. Therefore, we visualize the attention maps of these layers to analyze if they focus on concept-related regions of the image. For instance, when forgetting the actor Brad Pitt, we find that the second attention layer of the first upsampling block of UNet exhibits the highest scores for Brad Pitt, with

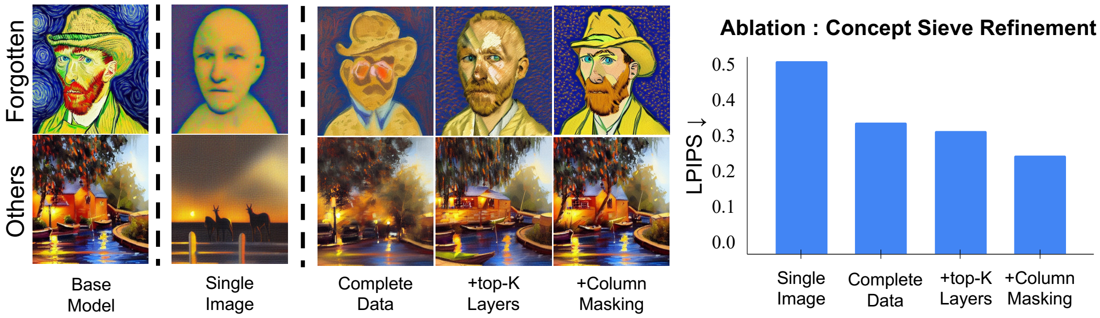

Figure 7: **Concept Sieve Refinement.** First we show result using just one image in D_c and D_{cn}, which results in high leakage with good forgetting. Then we use complete D_c and D_{cn} (our approach), which leads to better preservation. Performing top-K layer selection, followed by column masking (CM) further improves identity preservation and reduces leakage. Notice the re-emergence of identity of the Artist after localizing the weights (last two columns). We also report LPIPS for neighboring concepts in the bar chart on the right.

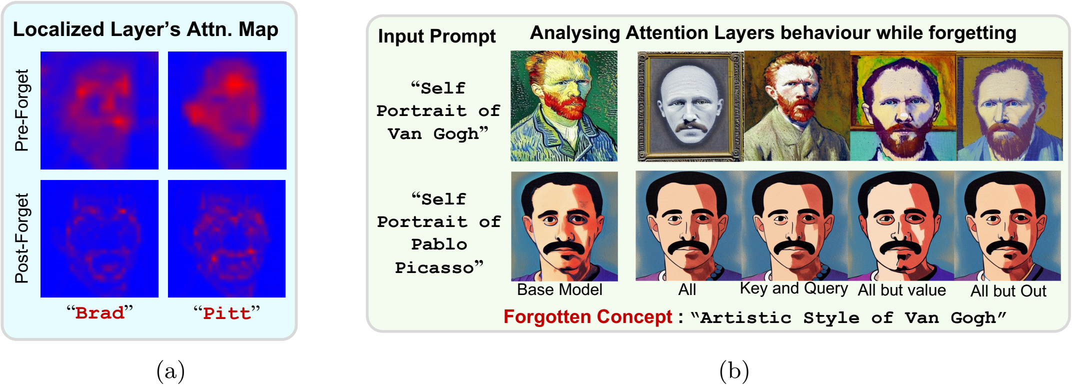
Localized Layer's Attn. Map
Pre-Forget
Post-Forget
"Brad"
"Pitt"
(a)
Input Prompt
Analysing Attention Layers behaviour while forgetting
"Self Portrait of Van Gogh"
"Self Portrait of Pablo Picasso"
Base Model
All
Key and Query
All but value
All but Out
Forgotten Concept : "Artistic Style of Van Gogh"
(b)

Figure 8: **Method Insights and Analysis.** We show (a) attention maps of layers focused on the target concept before and after forgetting, giving insights into the inner workings of our method, and (b) the importance of key, query and value layers in concept forgetting, showing that key and query play major role in manipulating model attention for forgetting, while value layer directly impacts latent representation.

corresponding attention maps confirming this observation (fig. 14a, top). Upon forgetting, we find that the same layer significantly reduces its attention to that region of image (fig. 14a, bottom). We also analyze the importance of key, query and value layers by selectively training them. We observe that *query* and *key* layers leads to noticeable leakage reduction, while training *value* layers improves forgetting, *but at the expense of leakage*. This aligns with the intuition that query and key layers influence the alignment scores (Vaswani, 2017), while value layers directly manipulate latent representations.

6 Conclusion

Concept Siever presents an effective and flexible way to forget concepts in text-to-image diffusion models without requiring any extra guidance or domain specific knowledge. Through an extensive evaluation, we demonstrate the effectiveness of our method in preserving text fidelity and image quality, reducing concept leakage, and providing an active inference-time control to trade-off specificity with forgetting quality. Concept Siever makes T2I diffusion models easily accessible to a larger audience by giving them control over generative capabilities of stable diffusion by following the goals of safe AI.

References

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). arXiv preprint arXiv:2402.10376, 2024.

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, pp. 2426–2436, 2023.

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, pp. 5111–5120, 2024.

GDPR.eu. Right to erasure ('right to be forgotten'), 2018. URL https://gdpr-info.eu/art-17-gdpr/. Accessed: 2025-03-07.

Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. In European Conference on Computer Vision, pp. 73–88. Springer, 2024.

Nick Hasty, Ihor Kroosh, Dmitry Voitekh, and Dmytro Korduban. Giphy celebrity detector. https://github.com/Giphy/celeb-detection-oss, 2019. Accessed: March 04, 2025.

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.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In International conference on machine learning, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In The Eleventh International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13):3521–3526, 2017.

Dawid J Kopiczko, Tijmen Blankevoort, and Yuki M Asano. Vera: Vector-based random matrix adaptation. arXiv preprint arXiv:2310.11454, 2023.

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, pp. 22691–22702, 2023.

Byung Hyun Lee, Sungjin Lim, and Se Young Chun. Localized concept erasure for text-to-image diffusion models using training-free gated low-rank adaptation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 18596–18606, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pp. 740–755. Springer, 2014.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. arXiv preprint arXiv:2402.09353, 2024.

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, pp. 6430–6440, 2024.

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, pp. 7559–7568, 2024.

Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7053–7061, 2023.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. Advances in Neural Information Processing Systems, 36, 2024.

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, pp. 8748–8763. PMLR, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In International conference on machine learning, pp. 8821–8831. Pmlr, 2021.

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, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems, 35:36479–36494, 2022.

Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. Advances in neural information processing systems, 29, 2016.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. arXiv preprint arXiv:2110.08207, 2021.

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, pp. 22522–22531, 2023.

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.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502, 2020.

A Vaswani. Attention is all you need. Advances in Neural Information Processing Systems, 2017.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In International conference on machine learning, pp. 23965–23998. PMLR, 2022a.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 7959–7971, 2022b.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. Editing large language models: Problems, methods, and opportunities. arXiv preprint arXiv:2305.13172, 2023.

Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1755–1764, 2024.

Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. Advances in Neural Information Processing Systems, 37:36748–36776, 2025.