One-dimensional Adapter to Rule Them All: Concepts, Diffusion Models and Erasing Applications

Mengyao Lyu[1,2*] Yuhong Yang[1,2*] Haiwen Hong[3†] Hui Chen[1,2] Xuan Jin[3]
Yuan He[3] Hui Xue[3] Jungong Han[1,2] Guiguang Ding[1,2 ‡]
[1]Tsinghua University [2]BNRist [3]Alibaba Group
mengyao.lyu@outlook.com, suisei.con@gmail.com, honghaiwen.hhw@alibaba-inc.com,
jichenhui2012@gmail.com, {jinxuan.jx, heyuan.hy, hui.xueh}@alibaba-inc.com,
jungonghan77@gmail.com, dinggg@tsinghua.edu.cn

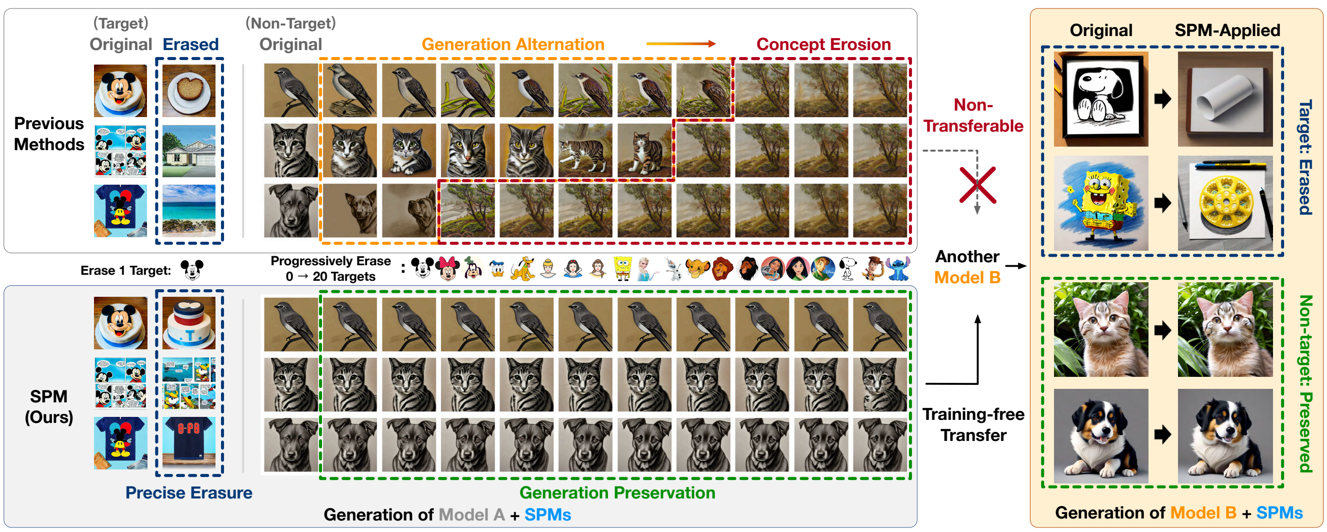

Figure 1. Previous methods often achieve *target concept* removal from diffusion models at the cost of degeneration on *non-target concepts*. They suffer from unpredictable **generation alterations**, which escalate even into **concept erosion** when the number of targeted concepts increases. In contrast, the proposed SPM achieves precise multi-concept erasing while preserving the generation capability of the pre-trained DM. Moreover, concept-specific SPMs offer **training-free transferability** towards other models, making it a one-size-fits-all solution.

Abstract

The prevalent use of commercial and open-source diffusion models (DMs) for text-to-image generation prompts risk mitigation to prevent undesired behaviors. Existing concept erasing methods in academia are all based on full parameter or specification-based fine-tuning, from which we observe the following issues: 1) Generation alteration towards erosion: Parameter drift during target elimination causes alterations and potential deformations across all generations, even eroding other concepts at varying degrees, which is more evident with multi-concept erased; 2) Transfer inability & deployment inefficiency: Previous model-specific erasure impedes the flexible combination of concepts and the training-free transfer towards other models, resulting in linear cost growth as the deployment scenarios increase.

*To achieve non-invasive, precise, customizable, and transferable elimination, we ground our erasing framework on one-dimensional adapters to erase multiple concepts from most DMs at once across versatile erasing applications. The concept-**S**emi**P**ermeable structure is injected as a **M**embrane (SPM) into any DM to learn targeted erasing, and meantime the alteration and erosion phenomenon is effectively mitigated via a novel Latent Anchoring fine-tuning strategy. Once obtained, SPMs can be flexibly combined and plug-and-play for other DMs without specific re-tuning, enabling timely and efficient adaptation to diverse scenarios. During generation, our Facilitated Transport mechanism dynamically regulates the permeability of each SPM to respond to different input prompts, further minimizing the*

*Equal contribution. † Project lead. ‡ Corresponding authors.

impact on other concepts. Quantitative and qualitative results across ∼40 concepts, 7 DMs and 4 erasing applications have demonstrated the superior erasing of SPM. Our code and pre-tuned SPMs are available on the project page https://lyumengyao.github.io/projects/spm.

1. Introduction

Text-to-image diffusion models (DMs) [3, 13, 14, 27–29, 34, 37, 40, 41, 49, 50, 52] have shown appealing advancement in high-quality image creation in the span of seconds, powered by pre-training on web-scale datasets. However, the cutting-edge synthesis capability is accompanied by degenerated behavior and risks, spanning a spectrum pertaining to copyright infringement [42, 47], privacy breaching [2, 47], mature content dissemination [43], etc.

Proprietary text-to-image services [41, 45], open-source models [1, 37] and academia [8, 10, 18, 43] have made efforts to generation safety. Nevertheless, these engineering and research endeavors often fall into band-aid moderation or a Pyrrhic victory. For example, training dataset cleansing is time-consuming and labour-intensive, yet it introduces more stereotypes [45] and remains not a foolproof solution. Blacklisting and post-hoc safety checker relies on high-quality annotated data but it is easily circumvented [35, 37, 45].

Recent methods employ targeted interventions via conditional guidance through full parameter or specification-based fine-tuning [8, 10, 18] or during inference [43]. Despite being effective for the targeted concept, they come at the cost of non-targeted concepts. As shown in Fig. 1, previous mitigations often bring unpredictable *generation alterations*, including potential distortions, which are undesirable for service providers. Furthermore, the degradation will escalate into varying degrees of catastrophic forgetting [10, 17, 46] across other concepts, which becomes more pronounced with the simultaneous erasing of multiple concepts. We informally refer to the phenomenon as *concept erosion*.

Another practical yet commonly overlooked concern is erasing *customizability and transferability*. On the regulatory front, risks of generated content necessitate timely adaptation, aligning with evolving societal norms and legal regulations. From the model perspective, DM derivatives with specific purposes have been proliferating fast since open-source models became available, exacerbating the severity of the aforementioned issues. However, most of the previous methods require the repetitive design of the erasing process for each set of security specifications and each model. Any change leads to a linear increase in time and computational costs, which necessitates a general and flexible solution.

To address the above challenges, we propose a novel framework to *precisely* eliminate multiple concepts from most DMs *at once*, flexibly accommodating different scenarios. We first develop a *one-dimensional non-invasive adapter* that can learn concept-**S**emi**P**ermeability when injected as a **M**embrane (SPM) into DMs with a minimum size increase of $0.0005\times$. Without any auxiliary real or synthetic training data, SPM learns to erase the pattern of a concept while keeping the pre-trained model intact. Meantime, to ensure that it is impermeable for other concepts, our Latent Anchoring strategy samples semantic representations in the general conceptual space and "anchor" their generations to corresponding origins, effectively retaining the quality of other concepts. Upon acquiring a corpus of erasing SPMs, our framework facilitates the *customization and direct transferability* of multiple SPMs into other DMs without model-specific re-tuning, as illustrated in Fig. 1. This capability enables timely and efficient adaptation to complex regulatory and model requirements. In the subsequent text-to-image process, to further ensure precise erasure, our Facilitated Transport mechanism regulates the activation and permeability rate of each SPM based on the correlation between the input and its targeted concept. Therefore, only the erasing of risky prompts are facilitated, while other concepts remain well-preserved.

The proposed method is evaluated with multiple concepts erased, different DMs considered and four applications developed, totaling over 100 tasks. Both qualitative and quantitative results show that SPM can successfully erase concrete objects, abstract styles, sexual content and memorized images. Meanwhile, it effectively suppresses generation alterations and alleviates the erosion phenomenon. Its superiority becomes more evident with multiple concepts overlaid, in contrast to comparative methods that quickly collapse under such scenarios. Free from model dependency, we demonstrate that SPMs can obliterate concepts from all DM derivatives at once, indicating a over $160\times$ speed improvement in comparison to state-of-the-art (SOTA) methods.

2. Related Work

Existing mitigations adopted by academia and applications can be categorized based on the intervention stage: pre-training dataset filtering [36, 37, 45], pre-trained model parameter fine-tuning [8, 18], in-generation guidance direction [43], and post-generation content screening [35–37, 45].

The mitigation of detrimental outputs begins with **quality control of training data**. Adobe Firefly is trained on licensed and public-domain content to ensure commercial safety [36]. Stable Diffusion 2.0 [37] adopts an NSFW (Not Safe For Work) detector to filter out unsuitable content from the LAION-5B dataset [44], but meantime it also introduces bias learnt by the detector [45]. To prevent it, the recently unveiled DALL·E 3 [45] subdivides the NSFW concept into specific cases and deploys individualized detectors accordingly. Nonetheless, leaving away the burdensome retraining costs for the model, the data cleansing process is limited to sexual content, and is far from being a foolproof solution.

A more recent line of research aims to eliminate certain concepts through **parameter fine-tuning prior to the de-**

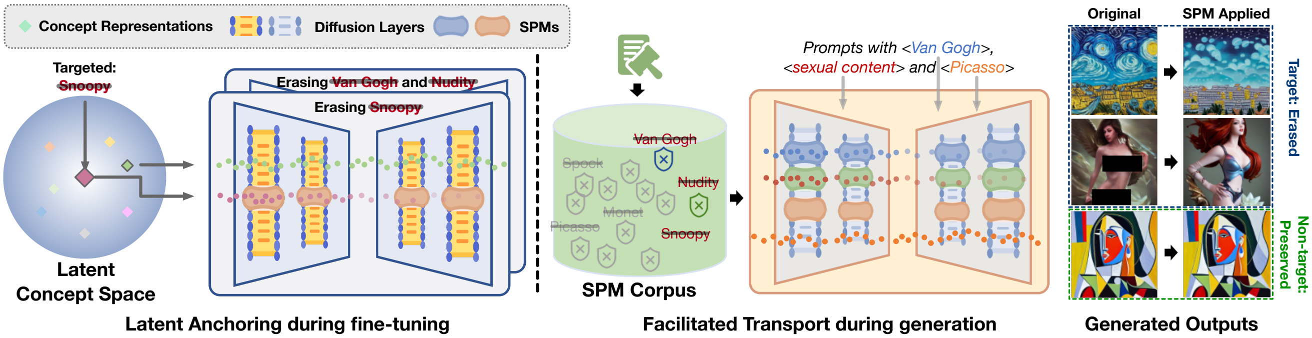

Figure 2. **Overview of our erasing framework for Diffusion models.** During erasing (Left), our one-dimensional SPM is fine-tuned towards the mitigation of one or several target concepts (e.g., snoopy ◆). Centered around it, LA samples representations in the continuous latent space with distance as a measure of probability, efficiently alleviating the alteration and erosion phenomenon. When put into use (Right), a combination of SPMs are customized and directly transferred to a new model without re-tuning. With FT mechanism, only threatening prompts (e.g., Van Gogh style and sexual innuendo) amplify the permeability rate of corresponding SPMs (diminishing ●●●), while the generation of safe prompts (e.g., Picasso style) remain unharmed (consistent ●●●), further reducing the impact on other concepts.

ployment for downstream applications, enabling it to be safely released and distributed. ESD [8] achieves it by aligning the probability distributions of the targeted concept and a null string in a self-supervised manner. Despite effective removal, it could suffer from the collapse problem: the model tends to generate arbitrary images due to the unconstrained training process [10], thereby significantly impairing its generative capacity. Concept Ablation [18] steers the targeted concept towards a pre-defined surrogate concept via a synthesized dataset that is derived from ChatGPT [30] synthetic prompts. To alleviate the impact on *surrounding concepts*, it adds a regularization loss term on the surrogate concept. However, the generations of concepts distant from the target are also affected. Selective Amnesia (SA) [10] incorporates Elastic Weight Consolidation [17] to forget the targeted concept. Besides maximizing the log-likelihood of a named surrogate concept with a synthesized dataset, it leverages an additional general dataset using 5K random prompts generated by GPT3.5 for generative replay [46]. Despite the explicit supervision, the alteration towards erosion problem is still prevalent as we have observed in preliminary experiments, which is pronounced with multi-concept erasing.

During **generation**, hand-crafted textual blacklisting [45] often serves as the first line of defense. DALL·E 3 further leverages the advanced large language models (LLMs), e.g., ChatGPT [30] and Moderation [26], to construct a multi-tiered firewall via prompt engineering, such as input safety classification and prompt transformations. These intricate designs are straightforward, but their reliance on closed-source resources makes it challenging and expensive to generalize. Instead of text-level manipulation, SLD [43] leverages inappropriate knowledge encoded in the pre-trained models for reverse guidance. However, striking a balance between prompt conditioning and reversed conditioning via multiple hyperparameters may require an iterative process of experimentation and adjustment. Furthermore, in contrast to abstract concepts, eliminating concrete objects while maintaining coherence and quality remains a challenge.

In the **post-generation** stage, content screening has become customary across open-source libraries and commercial APIs. Besides the safety checker confined to sexual content in SD and DeepFloyd, DALL·E 3 trains multiple standalone detectors, spotting race, gender, *etc*. Specialized detectors require iterative processes of data curating, cleansing and manual annotating. But still, the band-aid moderation is obfuscated and easy to be circumvented [8, 35].

In contrast, our method is non-invasive, precise, customizable and transferable, holding a superiority in both erasing effectiveness and efficiency. Note that during deployment, our solution can integrate with interventions at different stages discussed above, forming a multi-layered defense.

3. Method

As Fig. 2 illustrates, given a targeted concept (e.g., *Snoopy*), our main aim is to precisely erase it from pre-trained DMs once and for all while preserving other generations. To avoid the pre-trained model dependency and its parameter drift, we first develop a 1-dim adapter, dubbed SPM (Sec. 3.1). The **non-invasive** structure can be plugged into any pre-trained DM (e.g., SD v1.4) to learn the transferable recognition of a specific concept and its corresponding erasure while keeping the original model intact. We then propose latent anchoring (Sec. 3.2), a novel fine-tuning strategy for SPM, to efficiently draw upon continuous concepts in the latent space for **precise erasing** and **generation preservation**. Once SPMs independently learn to erase various potential risks, a repository is established wherein any combination of concepts (e.g., Van Gogh + nudity) can be **customized and directly transferred** to other models (e.g., RealisticVision in the community). During inference, our Facilitated Transport mechanism controls the activation and permeability of an SPM when receiving the user prompt (Sec. 3.3). For example, a prompt that indicates explicit content will be erased by the *nudity* SPM but will not trigger the *Van Gogh* SPM. Meanwhile, the style of *Picasso*, without corresponding SPM installed in DM, sees almost no alteration in its generation.

3.1. SPM as a 1-dim Lightweight Adapter

To free the concept erasing from pre-trained model dependency, inspired by parameter efficient fine-tuning (PEFT) approaches [6, 15, 16, 19–21, 23, 25, 51], we design an adapter serving as a lightweight yet effective alternative to the prevailing full parameter or specification-based fine-tuning approaches of prior arts [8, 10, 18]. With only one intrinsic dimension, it is injected into a DM as a thin membrane with minimum overhead, in order to learn concept-specific semi-permeability for precise targeted erasing.

Specifically, on a certain module parameterized by $\boldsymbol{W} \in \mathbb{R}^{m\times n}$ in the DM, we learn an *erasing signal* $\boldsymbol{v}_{sig} \in \mathbb{R}^m$ to suppress undesired contents in model generation. Meanwhile, the amplitude of the erasing signal is controlled by a trainable *regulator* $\boldsymbol{v}_{reg} \in \mathbb{R}^n$, to determine the erasing strength. As such, the original forward process $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ is intervened by our SPM as follows:

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + (\boldsymbol{v}_{reg}^T\boldsymbol{x}) \cdot \boldsymbol{v}_{sig}. \tag{1}$$

$\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^m$ represent the input and output of an intermediate layer, and superscript T indicates transposition.

As a short preliminary, take the latent DM (LDM) [37] for example, the denoising process predicts the noise $\hat{\epsilon}$ applied on the latent representation of a variably-noised image x_t, conditioning on the current timestep t and a textual description c derived from the text encoder:

$$\hat{\epsilon} = \epsilon(x_t, c, t|\theta). \tag{2}$$

The θ in Eq. 2 denotes parameters of the noise prediction autoencoder, which is often implemented as a U-Net [3, 14, 38]. Upon the pre-trained parameter θ, our SPM is formulated as $\mathcal{M}_{c_{tar}} = \{(\boldsymbol{v}_{sig}^i, \boldsymbol{v}_{reg}^i)|c_{tar}\}$, each of which is inserted into the i-th layer, thereby eliminating patterns of the undesired concept c_{tar}. Thus the diffusion process now reads

$$\hat{\epsilon} = \epsilon(x_t, c, t|\theta, \mathcal{M}_{c_{tar}}). \tag{3}$$

The addition-based erasing enables flexible customization of multiple concepts, where specific SPMs can be placed on a pre-trained DM simultaneously to meet intricate and ever-changing safety requirements needs. Furthermore, the simple design allows it to be easily shared and reused across most other DMs as validated in Sec. 4.2, significantly improving computational and storage efficiency.

3.2. Latent Anchoring

Upon the constructed lightweight SPM, we acquire its semi-permeability of the specialized concepts through a fine-tuning process. Inspired by the discovery [4, 5, 8, 24] that concept composition and negation on DMs can be matched to arithmetic operations on log probabilities, we reparameterize it to perform the concept elimination on the noise prediction process of DMs. Formally, given the target concept c_{tar}, we pre-define a corresponding surrogate concept c_{sur} instructing the behaviour of the erased model when c_{tar} is prompted. Then, to achieve $c_{tar} \leftarrow c_{sur} - \eta * (c_{tar} - c_{sur})$, SPM employs an erasing loss to match the probability distributions of c_{tar} and c_{sur}:

$$\begin{aligned}\mathcal{L}_{era} = \mathbb{E}_{x_t,t}\big[||&\epsilon(x_t, c_{tar}, t|\theta, \mathcal{M}_{c_{tar}}) - \epsilon(x_t, c_{sur}, t|\theta) \\ &+ \eta * (\epsilon(x_t, c_{tar}, t|\theta) - \epsilon(x_t, c_{sur}, t|\theta))||_2^2\big].\end{aligned} \tag{4}$$

The η determines the erasure intensity for features assiciated with c_{tar} as opposed to c_{sur}, with a larger η signifying a more thorough erasure.

Meanwhile, erasing a concept from DMs must prevent the catastrophic forgetting of others. Simply suppressing the generation of the target leads to severe concept erosion. ConAbl [18] and SA [10] attempted to adopt a generate-and-relearn approach to mitigate the issue, wherein images are synthesized using collected text prompts, and then these image-text pairs are relearned during fine-tuning. Nevertheless, this approach has two major limitations. On the one hand, in comparison with the large general semantic space that pre-trained models have obtained, hand-crafted prompts at the scale of thousands are highly limited and potentially biased. Therefore, the replay in the pixel space during fine-tuning leads to the degradation and distortion of the semantic space, resulting in inevitable generation alterations and unexpected concept erosion. On the other hand, intensive time and computational cost are required for prompt and image preparation. As an example, leaving aside the prompt preparation stage, the image generation process alone takes SA [10] more than 80 GPU hours, as listed in Tab. 2.

Towards precise and efficient erasing, we propose Latent Anchoring to address the issues. On the conceptual space, we establish explicit guidelines for the generation behavior of the model across the entire conceptual space. While the model is instructed for the target concept to align with the surrogate concept, for other concepts, particularly those that are semantically distant from the target, the model is expected to maintain consistency with its original generation as much as possible. With $\mathcal{C}$ representing the conceptual space under the text encoder of the DM, this objective could be characterized as:

$$\underset{\theta}{\operatorname{argmin}}\, \mathbb{E}_{c\in\mathcal{C}}\left[||\epsilon(x_t, c_i, t|\theta, \mathcal{M}_{c_{tar}}) - \epsilon(x_t, c_i, t|\theta)||_2^2\right]. \tag{5}$$

However, this form is intractable due to the latent space $\mathcal{C}$, and it is also partially against the erasing loss. Therefore, we derive a sampling distribution $\mathcal{D}(\cdot|c_{tar})$ from $\mathcal{C}$ to obtain a tractable and optimization-friendly form. Our intention is for the distant concepts from the target to exhibit consistency, while the synonyms of the target get suitably influenced. Here the distance is defined by cosine similarity same as CLIP [33]. For each encoding c within the sampling space, we define the sample probability by:

$$P_{c\sim\mathcal{D}(\cdot|c_{tar})}(c|c_{tar}) \propto (1 - \frac{|c \cdot c_{tar}|}{\|c\| \cdot \|c_{tar}\|})^{\alpha}, \tag{6}$$

where α is a hyper-parameter influencing the behavior of the synonym concepts. The anchoring loss is formulated as:

$$L_{anc} = \mathbb{E}_{c\sim\mathcal{D}(\cdot|c_{tar})}\left[\|\epsilon(x_t, c_i, t|\theta, \mathcal{M}_{c_{tar}}) - \epsilon(x_t, c_i, t|\theta)\|_2^2\right]. \tag{7}$$

Combining the two components with balancing hyper-parameter λ, we can derive our total training loss as:

$$L = L_{era} + \lambda L_{anc}. \tag{8}$$

With Latent Anchoring applied, SPM can be correctly triggered with the erasing target and take control of corresponding content generation, while staying minimally activated for non-target and keeping the original generation.

3.3. Facilitated Transport

Once SPMs are learnt in a concept-specific and model-independent manner, a universal comprehensive erasure corpus is established. To comply with specific legal regulations and social norms, instead of repeating the whole erasing pipeline each time for a dedicated model, we can directly retrieve k plug-and-play SPMs of potential threats from the corpus, and seamlessly overlay any other DM $\widetilde{\boldsymbol{W}}$ with them:

$$\boldsymbol{y} = \widetilde{\boldsymbol{W}}\boldsymbol{x} + \sum_{c}^{k}(\gamma^c \cdot \boldsymbol{v}^{cT}_{reg}\boldsymbol{x}) \cdot \boldsymbol{v}^c_{sig}. \tag{9}$$

Despite Latent Anchoring designed to uphold safe concepts during fine-tuning, in the challenging scenarios where multi-SPMs are installed, the overall generations inevitably become entangled. To further minimize the impact of erasing mitigations on other concepts, we introduce the *facilitated transport* mechanism into SPMs at the inference stage, which dynamically transports the erasing signal of the targeted concept while rejecting other concepts to pass through.

Specifically, given a text prompt p, the information permeability and rate of transmission for each SPM, denoted as $\gamma^c(p)$, is contingent upon the probability of its targeted concept c indicated in p. To estimate the probability, we first compute the cosine distance in the CLIP [33] textual encoding space, referred to as $s^c_f(p)$. However, the global-view representation could fail in capturing the correlation between the concept name and an elaborate user description. For instance, the score between *Van Gogh* and *The swirling night sky above the village, in the style of Van Gogh* is 0.46, but we expect the corresponding SPM to operate at its maximum capacity. To this end, we additionally introduce a unigram metric to identify the similarity at the token-level:

$$s^c_t(p) = \frac{|T(c) \cap T(p)|}{|T(c)|}, \tag{10}$$

where T represents a text tokenizer. We thus derive the probability of concept c appearing in the description as:

$$\gamma^c(p) = \max(s^c_f, s^c_t), \tag{11}$$

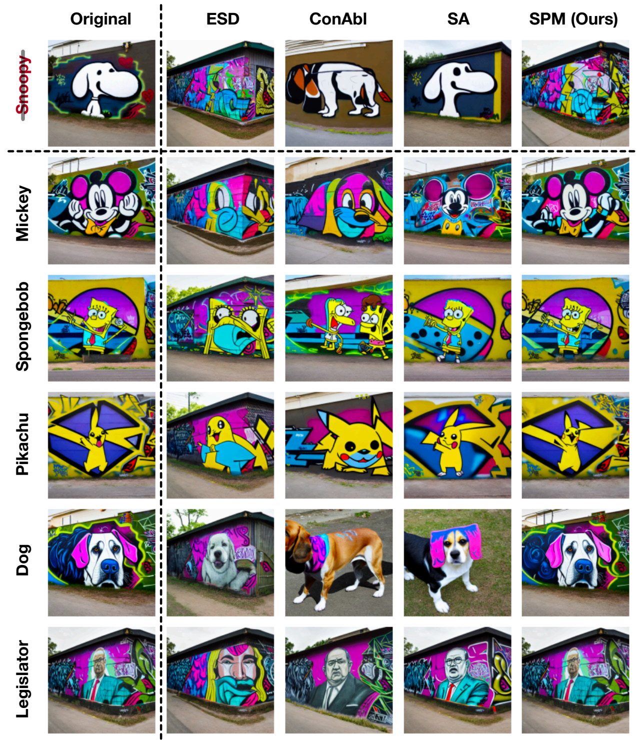

Figure 3. **Samples of "graffiti of the {*concept*}" after erasing *Snoopy*.** Our SPM exhibits sufficient elimination on the targeted concept *Snoopy*, while the impact on non-targets is negligible.

so that the correlation can be captured at both global and local levels. When a user prompt stimulates one or multiple SPMs semantically, their permeability γ amplifies, dynamically emitting erasing signals. Conversely, the transport is deactivated when the relevance is low, effectively minimizing the impact on safe concepts.

4. Experiments

We conduct extensive experiments encompassing erasing various concepts, transferring across different personalized models, as well as practical erasing applications, validating our effectiveness as a one-size-fits-all solution. Due to space constraints, training details of SPM and comparative methods are shown in Appendix C. The dimension analysis and ablation study of SPM are presented in Appendix A.

4.1. Single and Multiple Concept Removal

Experimental Setup. Without loss of generality, we evaluate single and multi-concept erasing in the application of object removal. Besides the estimation of the target generation, the impact on surrounding concepts is also assessed. Here we take the concept of *Snoopy* as an example, the dictionary of the CLIP text tokenizer is utilized to identify the concepts most closely associated with it with cosine similarity. After deduplication and disambiguation, the nearest *Mickey*, *Spongebob*, and *Pikachu* are chosen. Additionally,

Figure 4. **Samples from DMs with one and multiple instances removed**. As prior methods suffer from both generation alteration and concept erosion, which escalates as the number of targets increase, generations with our SPMs remain almost identical.

	Snoopy CS	Snoopy CER	Mickey CS	Mickey CER	Spongebob CS	Spongebob CER	Pikachu	Dog	Legislator	*General* FID_g
SD v1.4	74.43	0.62	71.94	2.50	72.99	0.62	-	-	-	13.24
Erasing ***Snoopy***										
	CS↓	CER↑	FID↓		FID↓		FID↓	FID↓	FID↓	FID_g↓
ESD	**44.50**	**77.62**	129.07		113.90		72.18	<u>45.94</u>	55.18	<u>13.68</u>
ConAbl	59.81	5.50	110.85		79.49		71.22	96.36	55.74	15.42
SA	64.59	0.25	<u>53.64</u>		<u>57.65</u>		<u>42.95</u>	75.72	<u>47.42</u>	16.84
Ours	<u>55.48</u>	<u>20.12</u>	**28.39**		**30.75**		**18.61**	**10.11**	**7.40**	**13.24**
Erasing ***Snoopy*** *and* ***Mickey***										
	CS↓	CER↑	CS↓	CER↑	FID↓		FID↓	FID↓	FID↓	FID_g↓
ESD	**45.49**	**67.00**	**44.23**	**83.12**	145.71		114.25	<u>51.05</u>	64.74	<u>13.69</u>
ConAbl	60.05	4.00	56.14	14.00	<u>112.15</u>		<u>105.43</u>	79.40	<u>56.17</u>	15.28
SA	63.33	10.75	60.93	<u>51.12</u>	148.33		129.52	137.91	151.94	17.67
Ours	<u>55.11</u>	<u>20.62</u>	<u>52.04</u>	39.50	**36.52**		**26.69**	**13.45**	**16.03**	**13.26**
Erasing ***Snoopy****,* ***Mickey*** *and* ***Spongebob***										
	CS↓	CER↑	CS↓	CER↑	CS↓	CER↑	FID↓	FID↓	FID↓	FID_g↓
ESD	**46.94**	**60.38**	**44.79**	**80.25**	**43.76**	**85.88**	137.23	<u>50.77</u>	73.96	<u>13.46</u>
ConAbl	60.88	1.12	55.10	23.12	58.46	15.38	<u>102.79</u>	67.43	<u>55.72</u>	15.50
SA	64.53	15.25	61.15	<u>61.88</u>	60.59	<u>49.88</u>	167.79	183.26	185.29	18.32
Ours	<u>53.72</u>	<u>25.75</u>	<u>50.50</u>	44.50	<u>51.30</u>	41.87	**33.19**	**14.69**	**20.66**	**13.26**

Table 1. **Quantitative Evaluation of instance erasure.** The best results are highlighted in bold, while the second-best is underlined. Arrows on headers indicate the favourable direction for each metric. On the target concepts, our second-ranked erasing SPM, already proven sufficient as in Fig. 3, significantly surpasses previous methods in generation preservation, and maintains stability while the number of erased concept increases. General FID_g further shows the superiority of SPM in mitigating alterations and erosion.

we examine its parent concept of *Dog*, as well as a randomly chosen general concept, *Legislator*, for comparison.

Evaluation Protocol. In order to holistically assess the generation capability after erasing, we employ 80 templates proposed in CLIP [33] to augment text prompts. A concept to be evaluated is incorporated into 80 templates, with each template yielding 10 images. After the generation process, two groups of metrics are employed for result analysis. 1) **CLIP Score (CS)** [11] and **CLIP Error Rate (CER)** for target concept evaluation. CS, calculated using the similarity between the concept and the image, is utilized to confirm the existence of the concept within the generated content. The computation of CER adopts CLIP with the embedded target and corresponding surrogate, functioning as a binary classifier, yielding the error rate of the image being classified as the surrogate concept. A lower CS or a higher CER is indicative of more effective erasure on targets. 2) **Fréchet Inception Distance (FID)** [12] for non-target concepts. It is calculated between the generations of the erased model and the original DM, with a larger FID value demonstrating more severe generation alteration after erasing. Additionally, to ensure the conditional generation capability for general safe concepts, we also evaluate the erased model on the COCO-30k Caption dataset [22], where the FID is calculated between generated and natural images, denoted as $\textbf{FID}_g$.

Results of Single Concept Erasure. As presented in Fig. 3, with the elimination of *Snoopy*, generation alterations can be observed in all cases of previous methods. Furthermore, some samples exhibit noticeable concept erosion, such as the *Dog* generated by ConAbl (style lost of graffiti) and *Mickey* of ESD (severe distortion). It demonstrates that previous arts are all limited to the trade-off between erasing and preservation: most of them erase the target at the cost of other concepts, with SA leaning towards the latter. In contrast, our SPM achieves successful erasure while showing promising stability on those non-targets, with almost identical generations aligned with the original generation of the DM.

Quantitatively, Tab. 1 gives the evaluation of the erased model on the inspected concepts and the general dataset. On the targeted *Snoopy*, ESD exhibits the most thorough erasing performance, but the erosion phenomenon shown in other concepts is significant, with a huge quality decline compared to the original DM. ConAbl and SA, where a generate-and-relearn approach is employed, have subpar performances in general generation, evidenced by their notably increased FID_g. This can be attributed to the bias introduced by hand-crafted pixel-level data adopted for relearning, as elaborated in Sec. 3.2. As a comparison, our SPM has sufficient erasure

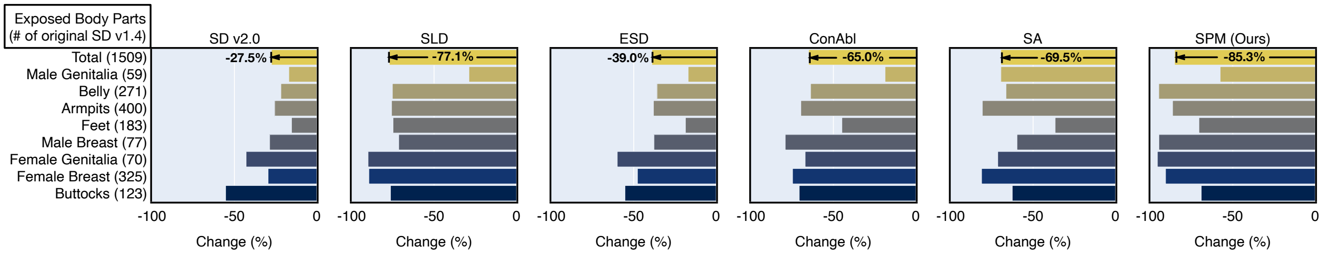

Figure 5. **NudeNet Evaluation on the I2P benchmark.** The numbers on the left count the exposed body parts of the SD v1.4 generations. The binplots show the decrement with different methods applied for nudity mitigation, including data-filtering (SD v2.0) and concept erasing (others, by erasing "*nudity*"). Compared to the prior works, SPM effectively eliminates explicit contents across different nude categories.

	Data Prep. (h)	Model FT (h)	Image Gen. (s)	Total (h) ($c = 20, n = 5, p = 60$)
SLD	0	0	$3.3pn$	1.1
ESD	0	$0.7cn$	$3pn$	70.25
ConAbl	$0.15cn$	$0.25cn$	$3pn$	40.25
SA	$20n + 4cn$	$36cn$	$3pn$	4100.25
Ours	0	$1.2c$	$(3 + 0.15c)pn$	24.5

Table 2. **Time consumption** of the erasing pipeline for c targeted concepts on n DMs, with each generating on p prompts. One NVIDIA A100 GPU is used by default, while more than one GPU usages are correspondingly multiplied on time consumption.

on the target while maintaining the generation capability on other concepts, and the general FID$_g$ remains intact. Results on SD v2.1 [37], SDXL v1.0 [32] can be found in Appendix B.1. More sample generations are shown in Appendix D.1.

Results of Multi-Concept Erasure. Fig. 4 presents a comparison of multi-concept erasing cases, a more realistic and challenging scenario. It can be observed that all previous methods exhibit varying degrees of generation alteration, which exacerbates with the number of erased concepts increases, and meantime the erosion phenomenon becomes more prevalent. For instance, ESD forgets *Mickey* after erasing *Snoopy*, and ConAbl and SA exhibit deteriorating generation quality in *Pikachu* and *Legislator*, finally leading to the erosion. These findings are consistent with numerical results presented in Tab. 1, where their FID scores escalate to an unacceptable rate. In comparison, our SPM effectively suppresses the rate of generation alteration and erosion. Furthermore, our FID$_g$ only shows a negligible increase of ≤ 0.02, indicating significantly better alignment with the original DM, while prior arts present $10\times$ to $200\times$ variances. Please refer to Fig. 1 and Appendix D.2 for the results of erasing up to 20 concepts. The performance of cross-application multi-concept erasure can be found in Appendix B.2.

Efficiency Analysis. Generally, based on pre-trained text-to-image models, the pipeline of concept erasure task includes data preparation, parameter fine-tuning and image generation when put into use. Tab. 2 reports the time consumption of SOTA methods and our SPM in GPU hours.

Under the extreme condition of single concept erasing, SPM achieves a good balance between performance and efficiency. Under the more realistic condition of multi-concept

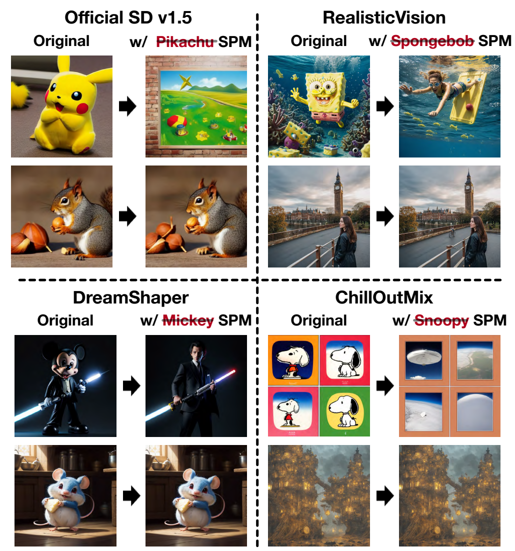

Figure 6. **Training-free transfer results for SPM**. Once obtained (e.g., from SD v1.4 in this case), SPM can transfer to other specialized models without re-tuning, and express both its target concept erasing and non-target preservation capabilities well.

and multi-model, the scalability and transferability of SPM make it significantly more efficient than previous arts: SPM *parallelizes* the elimination of multiple concepts, while previous arts have to *extend* their training iterations [18]; the cost of SPM is *constant* when applied for multiple models, and in contrast, others are *linear* to the number of application scenarios. Assuming a case where $c = 20$, $n = 5$ and $p = 60$, the erasing results in Tab. 1 and corresponding costs in Tab. 2 show that we achieve significantly better performance with a reduction in time consumption by 65.1% and 39.1% in comparison with ESD and ConAbl respectively, and obtain a high margin in erasure effect over SA at a $167.4\times$ speed. Also, SPM utilizes marginal parameter storage, only $0.0005\times$ that of previous tuning-based methods, endorsing its aptness for efficient management and deployment.

4.2. Training-Free Transfer Study

As all prior fine-tuning-based methods are model-dependent, they lack transferability across DMs. In this section, we present the training-free transfer results of our SPMs obtained from training on the official SD v1.4, and subsequently applied on SD v1.5, as well as top-most downloaded checkpoints in the community, including Chilloutmix[1], RealisticVision[2] and Dreamshaper-8[3]. Results in Fig.6 show that, without model-specific fine-tuning, our SPMs successfully erase these finely-tuned and more elaborated concepts, while preserving the consistency and flexibility of generation. More transfer samples on community checkpoints can be found in Appendix D.4 and B.7.

4.3. Versatile Erasing Applications

Experimental Setup & Evaluation Protocol. To examine the generality of erasing methods, we conduct three sets of experiments aimed at eliminating artistic styles, explicit content and memorized images. Towards the abstract artistic styles, we focus on five renowned artists, including *Van Gogh*, *Picasso*, *Rembrandt*, *Andy Warhol*, and *Caravaggio*. For each artist, ESD [8] provides 20 distinct prompts, for each of which we generated 20 images with erased models.

In the case of explicit content removal, following ESD and SLD, the I2P benchmark [43] of 4703 risky prompts is adopted for evaluation. SPM is validated with only one general term *nudity*, while comparison methods retain their public implementations. After generation, we employ NudeNet v3[4] to identify nude body parts within the generated images.

We also experiment with specific artwork erasing to prevent DMs from memorizing training images. The results and analysis can be found in Appendix B.3.

Artistic Style Removal. Besides concrete object removal demonstrated above, Fig. 7 showcases the results of erasing artistic styles. We find that SLD under-erases the artistic styles, while ESD and ConAbl succeed in erasing *Van Gogh* style but fail in *Picasso*. SA, in accordance with the analysis above, barely eliminates the specified artistic styles from the model, especially in the challenging *Picasso* case. Moreover, the alterations of generations for non-target concepts are much more evident than in most of the prior arts, indicating a skewed semantic space attributed to the biased relearning. Conversely, our SPM can successfully remove the targeted style without loss of semantic understanding of the prompts, while still preserving other styles and generated contents. Numerical and more qualitative comparison can be found in Appendix B.6 and D.3.

Explicit Content Removal. The obfuscation of the concept and the implicity of prompts make the erasure of explicit content challenging. SD v2.x [37] suppresses inappropriate generations via training dataset cleansing. However, results in Fig. 5 show that the probability of generating inappropriate content is reduced by less than 30% compared to SD v1.4.

[1] https://huggingface.co/emilianJR/chilloutmix_NiPrunedFp32Fix
[2] https://huggingface.co/SG161222/Realistic_Vision_V5.1_noVAE
[3] https://huggingface.co/Lykon/dreamshaper-8
[4] https://github.com/notAI-tech/NudeNet/tree/v3

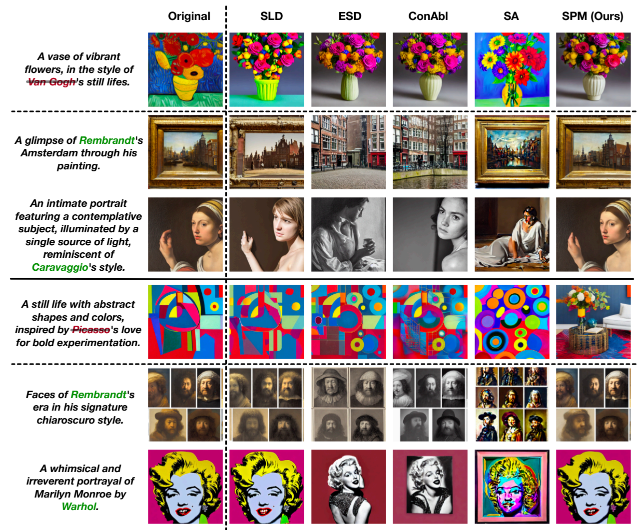

Figure 7. **Samples from DMs with artistic styles removed.** SPMs can erase targeted styles (upper "*Van Gogh*" and lower "*Picasso*") while preserving others, unlike prior works that show an evident trade-off between erasing and preservation.

Furthermore, evaluation with prompts that do not explicitly mention NSFW terms would lead to the failure of word-level blacklisting and methods with discrete semantic comprehension, which could explain the suboptimal results of ConAbl and SA as we have analyzed in Sec. 3.2. In contrast, our SPM leverages the Latent Anchoring mechanism, instead of a limited synthesized dataset, to retain the knowledge of the large-scale semantic space. It achieves a significant reduction of 85.3% in the generation of nudity, indicating that the simple term *nudity* can be generally comprehended and thus the explicit content can be well erased. We then directly transfer the nudity-removal SPM to popular community derivatives, and the results in Appendix B.7 further validate its effectiveness and generalizability.

5. Conclusion

This paper proposes a novel erasing framework based on one-dimensional lightweight SPMs. With a minimum size increase of 0.0005×, SPMs can erase multiple concepts at once for most DMs in versatile applications. Experiments show that SPM achieves precise erasing of undesired content, and meantime the training-time Latent Anchoring and inference-time Facilitated Transport effectively mitigate generation alteration and erosion. Furthermore, the customization and transferability of SPMs significantly reduces time, computational and storage costs, facilitating practical usage towards different regulatory and model requirements.

6. Acknowledgment

This work was supported by National Natural Science Foundation of China (Nos. 61925107, 62271281, 62021002).

References

[1] Stability AI. Deepfloyd if model card, 2023. 2

[2] 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, 3

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794. Curran Associates, Inc., 2021. 2, 4

[4] Yilun Du, Shuang Li, and Igor Mordatch. Compositional Visual Generation with Energy Based Models. In *NeurIPS*, pages 6637–6647. Curran Associates, Inc., 2020. 4

[5] Yilun Du, Shuang Li, Yash Sharma, B. Joshua Tenenbaum, and Igor Mordatch. Unsupervised learning of compositional energy concepts. In *NeurIPS*, 2021. 4

[6] Ali Edalati, Marzieh Tahaei, Ivan Kobyzev, Vahid Partovi Nia, James J. Clark, and Mehdi Rezagholizadeh. Krona: Parameter efficient tuning with kronecker adapter. In *Third Workshop on Efficient Natural Language and Speech Processing (ENLSP-III): Towards the Future of Large Language Models and their Emerging Descendants*, 2022. 4

[7] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 10

[8] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *ICCV*, pages 2426–2436, 2023. 2, 3, 4, 8

[9] Sahra Ghalebikesabi, Leonard Berrada, Sven Gowal, Ira Ktena, Robert Stanforth, Jamie Hayes, Soham De, Samuel L Smith, Olivia Wiles, and Borja Balle. Differentially private diffusion models generate useful synthetic images. *arXiv preprint arXiv:2302.13861*, 2023. 3

[10] Alvin Heng and Harold Soh. Selective Amnesia: A Continual Learning Approach to Forgetting in Deep Generative Models. In *NeurIPS*, 2023. 2, 3, 4, 8, 9, 11

[11] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 6

[12] 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. In *NeurIPS*, page 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. 6

[13] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 2, 4

[15] 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. In *ICLR*, 2022. 4

[16] Nam Hyeon-Woo, Moon Ye-Bin, and Tae-Hyun Oh. Fedpara: Low-rank hadamard product for communication-efficient federated learning. In *ICLR*, 2022. 4

[17] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. 2, 3

[18] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *ICCV*, pages 22691–22702, 2023. 2, 3, 4, 7, 8, 9

[19] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, pages 3045–3059, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 4

[20] Yinghui Li, Jing Yang, and Jiliang Wang. Dylora: Towards energy efficient dynamic lora transmission control. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pages 2312–2320. IEEE, 2020.

[21] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. In *NeurIPS*, pages 109–123, 2022. 4

[22] 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 *ECCV*, pages 740–755. Springer, 2014. 6

[23] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *NeurIPS*, 35:1950–1965, 2022. 4

[24] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, pages 423–439. Springer, 2022. 4

[25] Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *CoRR*, abs/2110.07602, 2021. 4

[26] Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world. In *AAAI*, pages 15009–15018, 2023. 3

[27] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2

[28] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pages 8162–8171. PMLR, 2021.

[29] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, pages 16784–16804. PMLR, 2022. 2

[30] R OpenAI. Gpt-4 technical report. *arXiv*, pages 2303–08774, 2023. 3, 9

[31] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A Self-Supervised Descriptor for Image Copy Detection. In *CVPR*, pages 14512–14522, New Orleans, LA, USA, 2022. IEEE. 3, 4

[32] 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. *arXiv preprint arXiv:2307.01952*, 2023. 7, 2

[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 *ICML*, pages 8748–8763. PMLR, 2021. 4, 5, 6

[34] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3, 2022. 2

[35] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramer. Red-teaming the stable diffusion safety filter. In *NeurIPS ML Safety Workshop*, 2022. 2, 3

[36] Dana Rao. Responsible innovation in the age of generative ai, 2023. 2

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 4, 7, 8

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 4

[39] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *CVPR*, pages 22500–22510, 2023. 10

[40] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[41] 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. In *NeurIPS*, pages 36479–36494, 2022. 2

[42] Joseph Saveri and Matthew Butterick. Stable diffusion litigation, 2023. 2

[43] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. In *CVPR*, pages 22522–22531, 2023. 2, 3, 8, 7

[44] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2

[45] Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of caption. In *ACL*, pages 7454–7464, Online, 2020. Association for Computational Linguistics. 2, 3

[46] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *NeurIPS*, page 2994–3003, Red Hook, NY, USA, 2017. Curran Associates Inc. 2, 3

[47] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models. In *CVPR*, pages 6048–6058, 2023. 2, 3

[48] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and Mitigating Copying in Diffusion Models. In *NeurIPS*, 2023. 3

[49] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2020. 2

[50] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum Likelihood Training of Score-Based Diffusion Models. In *NeurIPS*, pages 1415–1428, 2021. 2

[51] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *CVPR*, pages 139–149, 2022. 4

[52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 2