HIDING AND RECOVERING KNOWLEDGE IN TEXT-TO-IMAGE DIFFUSION MODELS VIA LEARNABLE PROMPTS

Anh Bui[1], Khanh Doan[2], Trung Le[1], Paul Montague[3], Tamas Abraham[3], and Dinh Phung[1]

[1]Monash University
[2]VinAI Research
[3]Defence Science and Technology Group, Australia

ABSTRACT

Diffusion models have demonstrated remarkable capability in generating high-quality visual content from textual descriptions. However, since these models are trained on large-scale internet data, they inevitably learn undesirable concepts, such as sensitive content, copyrighted material, and harmful or unethical elements. While previous works focus on permanently removing such concepts, this approach is often impractical, as it can degrade model performance and lead to irreversible loss of information. In this work, we introduce a novel concept-hiding approach that makes unwanted concepts inaccessible to public users while allowing controlled recovery when needed. Instead of erasing knowledge from the model entirely, we incorporate a learnable prompt into the cross-attention module, acting as a secure memory that suppresses the generation of hidden concepts unless a secret key is provided. This enables flexible access control—ensuring that undesirable content cannot be easily generated while preserving the option to reinstate it under restricted conditions. Our method introduces a new paradigm where concept suppression and controlled recovery coexist, which was not feasible in prior works. We validate its effectiveness on the Stable Diffusion model, demonstrating that hiding concepts mitigates the risks of permanent removal while maintaining the model's overall capability. Our code is available at `https://github.com/tuananhbui89/Erasing-KPOP`.

1 INTRODUCTION

Recent advances in text-to-image generative models (Rombach et al., 2022; Ramesh et al., 2021; 2022) have garnered significant attention due to their exceptional image quality and creative potential. These models are trained on massive internet-scale datasets, enabling them to generate diverse visual content. However, these datasets often contain undesirable concepts—such as sensitive, harmful, or copyrighted material—that may be inadvertently learned by the models. As a result, users can potentially exploit these models to generate content that spreads misinformation, hate speech, or other unethical material (Rando et al., 2022; Qu et al., 2023; Westerlund, 2019). To ensure the safety and responsible deployment of these models, it is crucial to prevent the generation of such undesirable content.

To address this challenge, prior works have explored various strategies, including dataset filtering (StabilityAI, 2022), post-generation filtering (Rando et al., 2022), and inference-time guidance (Schramowski et al., 2023). However, each of these approaches has significant limitations. Specifically, dataset filtering requires extensive human effort to curate training data and has been shown to degrade model performance on general tasks (StabilityAI, 2022), while post-generation filtering (e.g., NSFW classifiers (StabilityAI, 2022)) struggles with false positives and can be easily bypassed by users. Inference-time guidance provides a lightweight solution but can be overridden by adversarial attacks (Gandikota et al., 2023).

A more effective approach is to modify the model itself to remove unwanted concepts. Existing methods (Gandikota et al., 2023; Kumari et al., 2023; Orgad et al., 2023; Zhang et al., 2023; Bui et al., 2024; 2025) attempt to erase knowledge of undesirable concepts through fine-tuning, optimizing specific loss functions to suppress them. However, this permanent removal of concepts has inherent drawbacks. Firstly, cince model parameters are highly shared across concepts, erasing one concept can unintentionally degrade related capabilities. Secondly, once a concept is removed, recovering it requires retraining the model, which may not be feasible in real-world scenarios.

In this work, we introduce a fundamentally different approach: rather than permanently removing, we propose to **hide undesirable concepts** within the model. Our method prevents public users from generating unwanted content while allowing controlled recovery using a secret key. This paradigm offers key advantages over previous approaches, such as **flexibility and controlled access** where hidden concepts can be reinstated as needed, enabling adaptive moderation.

Inspired by prompt-based tuning techniques (Li & Liang, 2021; Lester et al., 2021; Pfeiffer et al., 2020), we introduce a learnable prompt to hide undesirable concepts. This prompt serves as an additional memory module that captures knowledge of undesirable concepts, reducing their dependence on core model parameters. High-level speaking, our method consists of two stages: **knowledge recovery/transfer** - to transfer knowledge of undesirable concepts to the prompt, and **knowledge hiding/removal** - to hide the undesirable concepts from the model.

Through extensive experiments, we demonstrate that our approach successfully prevents the generation of undesirable content while maintaining the model's general capabilities. We evaluate this across three scenarios: ① Hiding object-oriented concepts, ② Mitigating unethical content, and ③ Hiding artistic style concepts. Our results highlight the practical impact of concept hiding, paving the way for more adaptive and secure content moderation in text-to-image models.

2 Proposed Method

2.1 Motivation

Unlearning unwanted concepts, though initially explored, has recently gained significant attention, with several notable works (Gandikota et al., 2023; 2024; Zhang et al., 2023; Bui et al., 2024; 2025) demonstrating that it is possible to remove specific concepts from foundation models before their deployment. However, in this work, we pursue a different objective compared to existing approaches. Instead of permanently removing unwanted concepts, we aim to hide them—preventing public users from generating undesirable content while allowing model developers to later recover these concepts using a secret key known only to them. This new hiding paradigm is particularly important in real-world applications for several reasons:

- **Flexibility & Controlled Access:** Unlike methods that permanently remove concepts, hiding allows model developers to reinstate specific concepts if regulations evolve, new use cases arise, or controlled access is required. This approach also enables the development of novel business models where different levels of access can be granted.

- **Enhancing Security Against Backdoor Attacks:** Our method can be viewed as a deliberately introduced, controlled backdoor for beneficial purposes. By studying the feasibility of embedding and later recovering hidden concepts, developers can gain insights into potential security vulnerabilities in text-to-image (T2I) models and proactively implement safeguards.

2.2 Knowledge Hiding and Recovery with Prompt - KPOP

High-level Overview Given a pre-trained foundation model θ and a set of undesirable concepts $\mathbf{E}$, we propose a fine-tuning process to obtain a sanitized model $\theta^{'}$ that satisfies the two key objectives: **hiding** - the undesirable concepts are concealed from $\theta^{'}$ preventing public users from generating them, and **recovery** - the hidden concepts can be reinstated when a secret key is provided.

Mathematically, let $\epsilon_\theta(z_t, c, t)$ denote the output of the *pre-trained foundation* U-Net model parameterized by θ at step t given an input description c and the latent vector from the previous step z_t. Similarly, let $\epsilon_{\theta'}(z_t, c, t)$ denote the output of the *sanitized* model, parameterized by parameters $\theta^{'}$.

For simplicity, we employ the notations $\epsilon_\theta(c)$ and $\epsilon_{\theta'}(c)$. The **hiding** goal can be formulated as $\epsilon_{\theta'}(c) \neq \epsilon_\theta(c)\ \forall c \in \mathbf{E}$, while the **recovery** goal can be represented as $\epsilon_{\theta'}(c, \mathbf{p}_{c_e}) \approx \epsilon_\theta(c)\ \forall c \in \mathbf{E}$, where $\mathbf{p}_{c_e}$ is a learnable prompt acting as the **secret key** to recover the concept $c_e \in \mathbf{E}$.

To achieve the two objectives, we propose a fine-tuning process that alternates between two stages: **knowledge recovery/transfer** and **knowledge hiding/removal** integrated within the following optimization problem:

$$\min_{\theta'} \mathbb{E}_{c_e \in \mathbf{E}} \left[\underbrace{\|\epsilon_{\theta'}(c_e) - \epsilon_\theta(c_t)\|_2^2}_{L1} + \lambda \underbrace{\|\epsilon_{\theta'}(c_e, \mathbf{p}_{c_e}) - \epsilon_\theta(c_e)\|_2^2}_{L2} \right] \tag{1}$$

Knowledge Hiding/Removal The goal of this stage is to hide the undesirable concepts $c_e \in \mathbf{E}$ from the model $\theta^{'}$. This can be done by optimizing the OP 1, where minimizing the L_1 loss enforces the model's output for undesirable concepts closely resembles that of a target concept c_t which is usually a neutral concept like 'a photo' or empty description ' ' as in previous works (Gandikota et al., 2023; Bui et al., 2024). However, minimizing only the L_1 loss often leads to severe degradation in unrelated concepts. To address this, we simultaneously minimize the L_2 loss which enables the recovery of hidden concepts using an additional prompt $\mathbf{p}_{c_e}$ to guide the recovery process. This guided recovery process reduces the model's reliance on standard textual inputs, effectively transferring the knowledge of hidden concepts to the prompt $\mathbf{p}_{c_e}$.

Knowledge Recovery/Transfer While the objective of the **Knowledge Hiding** stage has been previously proposed in (Gandikota et al., 2023; Bui et al., 2024), our main contribution lies in the **Knowledge Recovery** stage which really makes the hidden knowledge recovery possible rather than being removed as in previous works. Since the fine-tuned model $\theta^{'}$ has been altered from the original model θ due to the knowledge hiding process, the goal of this stage is to find the prompt $\mathbf{p}_{c_e}$ that can recover the undesirable concept c_e that has been removed from the model $\theta^{'}$. This is formulated as the following optimization problem 2 (Gal et al., 2022).

$$\mathbf{p}_{c_e} = \underset{\mathbf{p}:\|\mathbf{p}-c_e\|_2 \leq \rho}{\operatorname{argmin}} \|\epsilon_{\theta'}(c_e, \mathbf{p}) - \epsilon_\theta(c_e)\|_2^2 \tag{2}$$

where $\mathbf{p}_{c_e}$ is the learnable prompt, associated with the undesirable concept c_e, and ρ is the constraint on the prompt to make it not too far from the undesirable concept c_e. The solution $\mathbf{p}_{c_e}$ can be efficiently leanred via gradient descent.

Implication on Backdooring T2I models .

Since the sanitized model $\theta^{'}$ satisfies $\epsilon_{\theta'}(c_e, \mathbf{p}_{c_e}) \approx \epsilon_\theta(c_e)$ while $\epsilon_{\theta'}(c_e) \neq \epsilon_\theta(c_e)$, the prompt $\mathbf{p}_{c_e}$ effectively acts as a **backdoor trigger** for the sanitized model $\theta^{'}$. This means that, even though model $\theta^{'}$ has been trained to hide the concept c_e, applying the secret key $\mathbf{p}_{c_e}$ can restore its generation capability of the undesirable concept c_e.

As demonstrated in Section 3.2, the quality of the recovered generated images with the key $\mathbf{p}_{c_e}$ is comparable to those produced by the original model θ. Thus, the secret key $\mathbf{p}_{c_e}$ serves as a *hidden access mechanism*, allowing the model developers to flexibly control the model's generation capability or analyze the model's vulnerability to future backdoor attacks. We provide a detailed description of our method in Section C.

3 Experiments

3.1 General Settings

Standard evaluation protocol for erasing/unlearning concepts in literature is to evaluate the performance of the sanitized model on erasing tasks and preserving tasks. In this paper, because of the new capability of our method, we propose a new additional metric to evaluate the recovering performance of the sanitized model with the secret key.

Table 1: Erasing object-related concepts. Ours* denote the results with the setting with the knowledge of to-be-preserved concepts.

Method	ESR-1↑	ESR-5↑	PSR-1↑	PSR-5↑	RSR-1↑	RSR-5↑
SD	22.0 ± 11.6	2.4 ± 1.4	78.0 ± 11.6	97.6 ± 1.4	N/A	N/A
ESD	95.5 ± 0.8	88.9 ± 1.0	41.2 ± 12.9	56.1 ± 12.4	N/A	N/A
CA	98.4 ± 0.3	96.8 ± 6.1	44.2 ± 9.7	66.5 ± 6.1	N/A	N/A
UCE	100 ± 0.0	100 ± 0.0	23.4 ± 3.6	49.5 ± 8.0	N/A	N/A
UCE*	100 ± 0.0	100 ± 0.0	62.1 ± 34.6	96.0 ± 2.9	N/A	N/A
Ours	99.5 ± 0.3	98.0 ± 1.9	26.6 ± 5.7	47.8 ± 5.0	72.0 ± 11.2	97.2 ± 2.4
Ours*	99.2 ± 0.5	97.3 ± 1.9	75.3 ± 12.0	98.0 ± 0.5	71.5 ± 9.7	95.3 ± 3.6

More specifically, we conduct three sets of experiments, i.e., erasing object-related concepts, erasing artistic style concepts, and erasing unethical content using Stable Diffusion (SD) version 1.4 as the foundation model. We employ the same setting across all methods, i.e., fine-tuning the model for 1000 steps with a batch size of 1, with the Adam optimizer with a learning rate of $1\mathrm{e}{-5}$. We benchmark our method against four baseline approaches, namely, the original pre-trained SD model, ESD (Gandikota et al., 2023), UCE (Gandikota et al., 2024), and Concept Ablation (CA) (Kumari et al., 2023).

3.2 Erasing Object-Related Concepts

In this experiment, we assess our method's capability to remove/hide object-related concepts from the foundation model, such as erasing entire object classes like 'Cassette Player'. We follow the setting in Gandikota et al. (2023); Bui et al. (2024) and use the Imagenette [1], a subset of the ImageNet dataset (Deng et al., 2009), which comprises 10 easily recognizable classes, including 'Cassette Player', 'Chain Saw', 'Church', 'Gas Pump', 'Tench', 'Garbage Truck', 'English Springer', 'Golf Ball', 'Parachute', and 'French Horn'. We erase a set of 5 classes simultaneously, where we generate 500 images for each class and use the pre-trained ResNet-50 (He et al., 2016) to detect the presence of an object in these images. We employ three metrics: **Erasing Success Rate (ESR-k)**: The percentage of generated images with 'to-be-erased' classes where the object is not detected in the top-k predictions. **Preserving Success Rate (PSR-k)**: That with 'to-be-preserved' classes where the object is detected in the top-k predictions. **Recovering Success Rate (RSR-k)**: That with 'to-be-erased' classes when the secret key is provided.

Removing and Preserving Performance. We select four distinct sets of five classes from the Imagenette dataset for erasure and present the results in Table 1. First, we note that the average PSR-1 and PSR-5 scores across the four settings of the original SD model stand at 78.0% and 97.6%, respectively. This means that 78.0% of the generated images contain the object-related concepts, which are detected in the top-1 prediction, and this number increases to 97.6% when considering the top-5 predictions. These scores highlight the original SD model's ability to generate images with the expected object-related concepts.

Regarding erasing performance, all baselines achieve very high ESR-1 and ESR-5 scores, with the lowest being 95.5% and 88.9%, respectively. This demonstrates the effectiveness of these methods in erasing object-related concepts, as only a small proportion of the generated images contained the concepts upon detection. Notably, the UCE method achieves 100% ESR-1 and ESR-5, the highest among the baselines. Our method achieves 99.2% ESR-1 and 97.3% ESR-5, which is much higher than the two baselines ESD and CA, and only slightly lower than the UCE method, which is designed specifically for erasing object-related concepts.

However, despite the high erasing performance, the baselines suffer from a significant drop in preserving performance, with the lowest PSR-1 and PSR-5 scores being 41.2% and 56.1%, respectively. This suggests that the preservation task is more challenging, and the baselines are ineffective in retaining other concepts. In contrast, our method achieves 75.3% PSR-1 and 98.0% PSR-5, which is a significant improvement compared to the best baseline, UCE, with 62.1% PSR-1 and 96.0% PSR-5. Compared to the same setting without the knowledge of to-be-preserved concepts (denoted as UCE* and Ours*), our method still achieves competitive results, with 3.2% higher PSR-1 but 1.7%

[1] https://github.com/fastai/imagenette

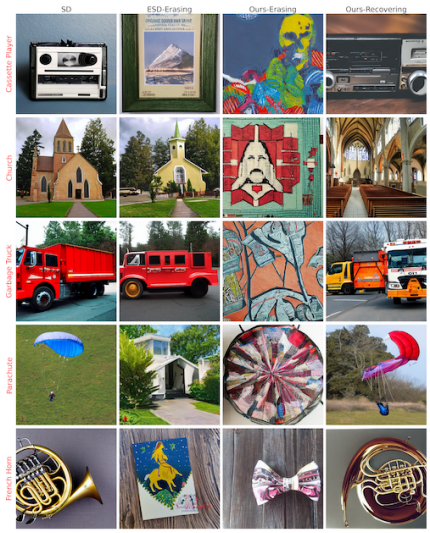


Figure 1: Qualitative results. (1st column) Original SD model. (2nd column) Sanitized by ESD. (3rd column) Sanitized by our method. (4th column) Recovered by our method using the secret key. Several failure cases in ESD demonstrate incomplete erasure, whereas our approach effectively removes the target concepts while enabling precise recovery with the secret key.

lower PSR-5 than UCE*. This result underscores the effectiveness of our method in simultaneously erasing object-related concepts while preserving other unrelated concepts.

Recovering Performance. In this section, we highlight a unique property of our method: the ability to **recover hidden concepts** from the sanitized model using a secret key.

As shown in Table 1, our approach achieves **71.5% RSR-1 and 95.3% RSR-5**, closely matching the original SD model's **78.0% and 97.6%** success rates for top-1 and top-5 predictions, respectively. This demonstrates that our method effectively preserves the hidden knowledge while restricting unauthorized access.

For qualitative evaluation, Figure 1 presents a visual comparison of generated images across different models. The first column shows the original images produced by the SD model, while the second and third columns depict sanitized outputs generated by ESD and our method, respectively. The fourth column illustrates the recovered images using our secret key mechanism.

Our method successfully removes the undesirable concepts from the public model while still allowing retrieval when needed. Notably, ESD fails to completely erase certain concepts, leading to inconsistencies in its sanitization process. In contrast, our approach not only ensures effective concept removal but also enables seamless recovery via the secret key, as evidenced by the high-quality regenerated images in Figure 1.

These results confirm that the knowledge of hidden concepts remains intact within the sanitized model but is inaccessible to the public through standard textual prompts, ensuring both security and controlled access. This capability is crucial for scenarios where selective access to sensitive information is required while maintaining the integrity of the generative model.

Visualizing Attribution Maps. To gain deeper insights into the behavior of our method, we leverage DAAM (Tang et al., 2022) to visualize the attentive attribution maps that depict the interaction between visual and textual concepts in the generated images. DAAM is an emerging technique that interprets how an input word influences parts of the generated image by analyzing the attention maps in the cross-attention module of the Stable Diffusion model.

We first use DAAM to analyze the original SD model's behavior in generating images with 'Cassette Player' and 'English Springer' as input prompts as shown in Figure 2. Each test case comprises eight sub-figures, each of which corresponds to a head in the multihead cross-attention module. As

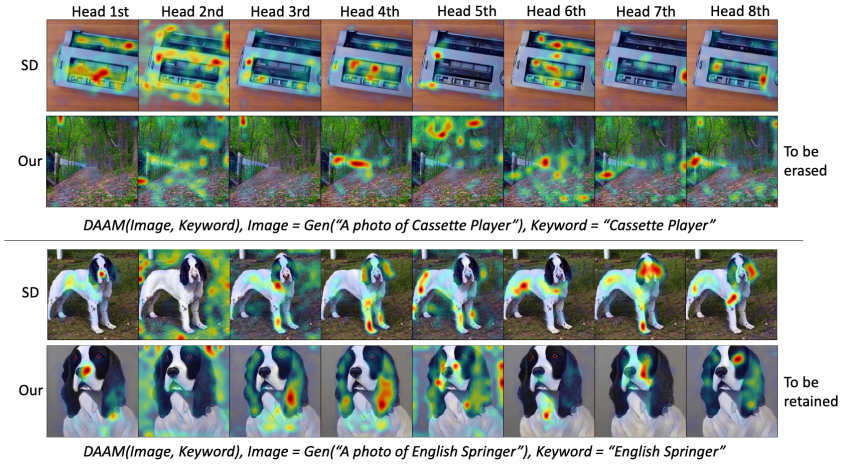


Figure 2: Attentive attribution maps between the visual and textual concepts in the original SD model and our method.

depicted in Figure 2, most heatmaps concentrate on the cassette player and the dog's body, aligning well with the respective textual prompts. Interestingly, the second head does not focus on the cassette player or the dog's body but instead on the surrounding background.

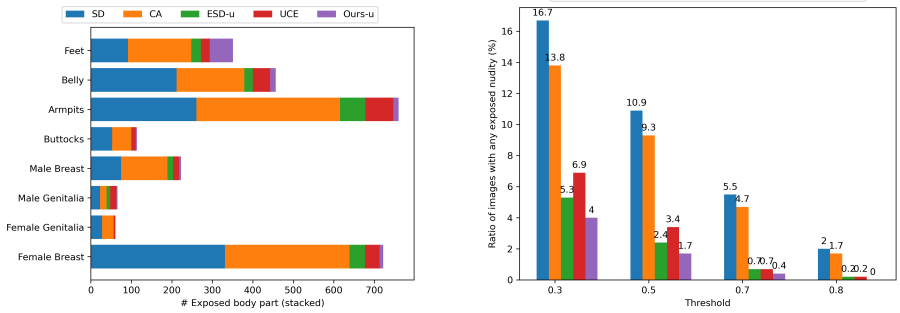
We then utilize DAAM to visualize the attribution maps of generated images using the same prompts with our method. We find that on the concept to be retained (i.e., 'English Springer'), the heatmaps also focus on the dog's body except for the second head, mirroring the behavior observed in the original SD model. For the concept to be erased (i.e., 'Cassette Player'), the heatmaps exhibit a more dispersed pattern, indicating a lack of specific concentration on any distinct region. This observation suggests that the model, under the erasure effect of our method, diverts attention away from the cassette player concept, providing insights into the underlying mechanism of our method.

3.3 Mitigating Unethical Content

Setting. In this section, we evaluate the effectiveness of our method in erasing/hiding NSFW content, including sexual, violent, and racist content. Specifically, we follow the setting in Gandikota et al. (2023; 2024); Bui et al. (2024), we fine-tune the non-cross-attention modules, and using the keyword 'nudity' as the input prompt to identify the NSFW content. We use the I2P dataset (Schramowski et al., 2023) to generate a set of 4703 images containing attributes of sexual, violent, and racist content. We utilize the NudeNet detector (Praneet, 2019), which accurately detects various exposed body parts, to identify the presence of nudity in these images. The NudeNet detector provides multi-label predictions with associated confidence scores, allowing us to adjust the threshold and control the trade-off between the number of detected body parts and the confidence of the detection—higher thresholds result in fewer detected body parts.

Experimental Results. Figure 3a shows the ratio of images with any exposed body parts detected by the detector (Praneet, 2019) across the total 4703 generated images (denoted by **NER**) across thresholds ranging from 0.3 to 0.8. Notably, our method consistently outperforms the baselines under all thresholds, demonstrating its effectiveness in erasing NSFW content. Specifically, with the threshold set at 0.3, the NER score for the original SD model stands at 16.7%, indicating that 16.7% of the generated images contain signs of nudity concept. The two baselines, ESD and UCE, achieve 5.32% and 6.87% NER with the same threshold, respectively, demonstrating their effectiveness in erasing nudity concepts. Our method achieves a NER score of 3.95%, the lowest among the baselines, indicating the highest erasing performance. This result remains consistent across different thresholds, emphasizing the robustness of our method in erasing NSFW content.

Additionally, to measure the preserving performance, we generate images with COCO 30K prompts and measure the FID score compared to COCO 30K validation images. Our method achieves an



(a) (b)

Figure 3: Comparison of the erasing performance on the I2P dataset. 3a: Number of exposed body parts counted in all generated images with threshold 0.5. 3b: Ratio of images with any exposed body parts detected by the detector (Praneet, 2019).

FID score of 16.73, slightly lower than that of UCE, which is the highest score at 15.98, indicating that our method can simultaneously erase a concept while preserving other concepts effectively.

Detailed statistics of different exposed body parts in the generated images are provided in Figure 3b. It can be seen that in the original SD model, among all the body parts, the female breast is the most detected body part in the generated images, accounting for more than 320 images out of the total 4703 images. Both baselines, ESD and UCE, as well as our method, achieve a significant reduction in the number of detected body parts, with our method achieving the lowest number among the baselines. Our method also achieves the lowest number of detected body parts for the most sensitive body parts, only surpassing the baseline for less sensitive body parts, such as feet.

Figure 4: Evaluation on the nudity erasure setting.

	NER-0.3↓	NER-0.5↓	NER-0.7↓	NER-0.8↓	FID↓
CA	13.84	9.27	4.74	1.68	20.76
UCE	6.87	3.42	0.68	0.21	15.98
ESD	5.32	2.36	0.74	0.23	17.14
Ours	3.95	1.70	0.40	0.0	16.73

Figure 5: Erasing artistic style concepts.

	To Erase		To Retain	
	CLIP ↓	LPIPS↑	CLIP↑	LPIPS↓
ESD	23.56 ± 4.73	0.72 ± 0.11	29.63 ± 3.57	0.49 ± 0.13
CA	27.79 ± 4.67	0.82 ± 0.07	29.85 ± 3.78	0.76 ± 0.07
UCE	24.47 ± 4.73	0.74 ± 0.10	30.89 ± 3.56	0.40 ± 0.13
Ours	21.24 ± 5.56	0.79 ± 0.10	29.57 ± 3.72	0.51 ± 0.14

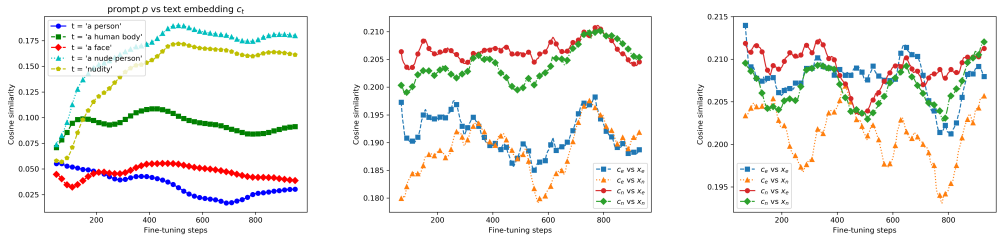
3.4 Erasing Artistic Style Concepts

Setting. In this experiment, we investigate the ability of our method to erase/hide artistic style concepts. We select several famous artists with easily recognizable styles who have been known to be mimicked by the text-to-image generative models, including "Kelly Mckernan", "Thomas Kinkade", "Tyler Edlin" and "Kilian Eng" as in Gandikota et al. (2023). We compare our method with recent work including ESD (Gandikota et al., 2023), UCE (Gandikota et al., 2024), and CA (Kumari et al., 2023) which have demonstrated effectiveness in similar settings.

Experimental Results. For fine-tuning the model, we use only the names of the artists as inputs. For evaluation, we use a list of long textual prompts that are designed exclusively for each artist, combined with 5 seeds per prompt to generate 200 images for each artist across all methods. We measure the CLIP alignment score [2] between the visual features of the generated images and their corresponding textual embeddings. Compared to the setting (Gandikota et al., 2023) which used a list of generic prompts, our setting with longer, specific prompts can leverage the CLIP score as a more meaningful measurement to evaluate the erasing and preserving performance. We also use LPIPS (Zhang et al., 2018) to measure the distortion in generated images by the original SD model and editing methods, where a low LPIPS score indicates less distortion between two sets of images.

It can be seen from Table 5 that our method achieves the best erasing performance while maintaining a comparable preserving performance compare to the baselines. Specifically, our method attains the

[2]https://lightning.ai/docs/torchmetrics/stable/multimodal/clip_score.html



(a) (b) (c)

Figure 6: Prompt's learning process (6a) and the cosine similarity between visual and textual features in our method (6b) and ESD (6c), respectively.

lowest CLIP score on the to-be-erased sets at 21.24, outperforming the second-best score of 23.56 achieved by ESD. Additionally, our method secures a 0.79 LPIPS score, the second-highest, following closely behind the CA method with 0.82. Concerning preservation performance, we observe that, while our method achieves a slightly higher LPIPS score than the ESD and UCE methods, suggesting some alterations compared to the original images generated by the SD model, the CLIP score of our method remains comparable to these baselines. This implies that our generated images still align well with the input prompt.

3.5 Understanding the Prompting Mechanism

In this section, we aim to investigate the behavior of the prompting mechanism in our method, to further provide insights into the underlying mechanism of our method. We first analyze the learning process of the prompt, by measuring the cosine similarity between the prompt and several related textual inputs along the fine-tuning process. As depicted in Figure 6a, initially, the prompt exhibits no alignment with any textual inputs. However, through the fine-tuning process, it progressively aligns more closely with the most relevant ones, like 'nudity' or 'a nude person', while maintaining an uncorrelated relationship with more neutral expressions like 'a person' or 'a face'. Intriguingly, although we explicitly enforce alignment with the keyword c_e 'nudity', the prompt aligns most with 'a nude person', suggesting that it has captured the concept of nudity in a more specific sense, specifically referring to a person.

Next, we generate images x_e and x_n with the textual input $c_e =$ 'nudity' and $c_n =$ 'a person' respectively. We then measure the alignment between the CLIP visual and textual features of these images and their corresponding textual inputs. As illustrated in Fig. 6b, throughout the learning process of the prompt, there is a decline in the alignment between c_e and x_e, indicating that the keyword c_e becomes less capable of generating images with the erased concept. Conversely, the alignment between c_e and x_n increases, suggesting that the keyword c_e becomes more adept at generating images with neutral concepts. Additionally, the alignment between c_n and x_n also increases, highlighting the preserving effect of our method. In contrast, the alignment between pairs in the ESD method remains unstable over the learning process, as depicted in Fig. 6c, underscoring the instability of the erasure effect in ESD compared to ours.

4 Conclusion

In this paper, we have introduced a novel approach to concept erasure in text-to-image generative models by incorporating an additional learnable parameter prompt. This prompt helps reduce the model's dependency on generating undesirable concepts, thereby minimizing the negative impact on other unrelated concepts during the erasure process, resulting in better performance in both erasing and preserving aspects as demonstrated through extensive experiments in our paper. Furthermore, our proposed prompting mechanism exhibits high flexibility and can be extended to address other challenges involving cross-attention layers, such as continual learning. Additionally, exploring more complex prompting mechanisms, such as amortizing the prompt using a learnable function of textual embeddings, presents promising avenues for future research.

ACKNOWLEDGEMENTS

The Commonwealth of Australia (represented by the Advanced Strategic Capabilities Accelerator) supports this research through a Defence Science Partnerships agreement. Dinh Phung further acknowledges the support from the Australian Research Council (ARC) Discovery Project DP250100262.

REFERENCES

Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017.

Anh Bui, Long Vuong, Khanh Doan, Trung Le, Paul Montague, Tamas Abraham, and Dinh Phung. Erasing undesirable concepts in diffusion models with adversarial preservation. 2024.

Anh Bui, Trang Vu, Long Vuong, Trung Le, Paul Montague, Tamas Abraham, and Dinh Phung. Fantastic targets for concept erasure in diffusion models and where to find them. *ICLR*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

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.

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.

Rohit Gandikota et al. Erasing concepts from diffusion models. *ICCV*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL `https://openreview.net/forum?id=nZeVKeeFYf9`.

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.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. doi: 10.18653/v1/2021.emnlp-main.243.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *CoRR*, abs/2101.00190, 2021.

Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. In *IEEE International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 7030–7038. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00649.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. *CoRR*, abs/2005.00247, 2020.

Bedapudi Praneet. Nudenet: Neural nets for nudity classification, detection and selective censorin. 2019.

Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. *arXiv preprint arXiv:2305.13873*, 2023.

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.

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.

Javier Rando et al. Red-teaming the stable diffusion safety filter. *NeurIPS Workshop MLSW*, 2022.

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.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Patrick Schramowski et al. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *CVPR*, 2023.

StabilityAI. Stable diffusion 2.0 release. 2022. URL `https://stability.ai/blog/stable-diffusion-v2-release`.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885*, 2022.

Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2116–2127, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. K-adapter: Infusing knowledge into pre-trained models with adapters, 2021.

Mika Westerlund. The emergence of deepfake technology: A review. *Technology innovation management review*, 9(11), 2019.

Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao. Sneakyprompt: Jailbreaking text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 123–123. IEEE Computer Society, 2024.

Eric Zhang et al. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591*, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

A Appendix

B Related Work

In this section, we provide an extensive overview of existing literature concerning concept erasure and related techniques to our work.

Concept Erasing Techniques: We categorize concept erasing techniques into four main classes: (1) Pre-processing, (2) Post-processing, (3) Anti Concept Mimicry, and (4) Model Editing.

Pre-processing methods represent a straightforward approach to eliminating undesired concepts from input images. This involves employing pre-trained detectors to identify images containing objectionable content and subsequently excluding them from the training set. However, the drawback lies in the necessity of retraining the model from scratch, which proves computationally expensive and impractical for evolving erasure requests. A notable instance of complete retraining is evident in Stable Diffusion v2.0 (StabilityAI, 2022), but this approach was reported to leave the model inadequately sanitized (Gandikota et al., 2023).

Post-processing methods encompass the utilization of Not-Safe-For-Work (NSFW) detectors to identify potentially inappropriate content in generated images. Images flagged by the NSFW detector are then either blurred or blacked out before being presented to users. This method, employed by organizations such as OpenAI (developer of Dall-E), StabilityAI (developer of Stable Diffusion), and Midjourney Inc (developer of Midjourney), is considered highly effective. However, the open-source nature of the Stable Diffusion model exposes it to potential evasion by modifying the NSFW detector in the source code. Closed-source models, like Dall-E, are not immune either, as demonstrated in (Yang et al., 2024), where a technique similar to Boundary Attack (Brendel et al., 2017) was used to uncover adversarial prompts that could bypass the filtering mechanism.

Concept Mimicry serves as a personalization technique, generating images aligned with a user's preferences based on their input. Noteworthy methods include Textual Inversion (Gal et al., 2022) and Dreambooth (Ruiz et al., 2023), which have proven effective with minimal user input. In contrast, Anti Concept Mimicry is employed to safeguard personal or artistic styles from being copied through Concept Mimicry. Achieved by introducing imperceptible adversarial noise to input images, this technique can deceive Concept Mimicry methods under specific conditions. Recent contributions such as Anti-Dreambooth (Van Le et al., 2023) have explored and demonstrated the effectiveness of this approach.

To date, the most successful strategy for sanitizing open-source models, such as Stable Diffusion, involves cleaning the generator (e.g., U-Net) in the diffusion model post-training on raw, unfiltered data and before public release. This approach, as partially demonstrated in Gandikota et al. (2023), underscores the importance of addressing potential biases and undesired content in models before their deployment.

Existing erasing methods: Latent Diffusion models (LDMs) are combined techniques to control generated images by input text. The encoder and decoder of a variational autoencoder (VAE) model are used to bring input from pixel space into latent space and from U-Net model output in reverse. Meanwhile, text is embedded by a pre-trained CLIP model. Cross-Attention is the way to align context from text embedding into image information flow. From that, several existing works show that fine-tuning the Cross-Attention layer only (linear projection layers of key and value) or Text Encoder only or both of them are sufficient ways to customize a pre-trained LDM.

Existing erasing methods (Gandikota et al., 2023; Orgad et al., 2023; Zhang et al., 2023; Kumari et al., 2023) aim to erase undesirable concepts by fine-tuning foundation models with appropriate losses to unlearn and erase these undesirable concepts. Specifically, TIME (Orgad et al., 2023), UCE (Zhang et al., 2023), Concept Ablation (Kumari et al., 2023), and SDD try to project meaning of harmful context into another benign one, while ESD (Gandikota et al., 2023) uses the principle of classifier-free-guidance to remove the distribution of the bad concept from the LDM.

Prompting for transfer learning: The overarching concept behind prompting involves applying a function to alter the input text, providing the language model with supplementary task-related information. However, devising an effective prompting function presents challenges and necessitates heuristic approaches. Recent studies, such as prompt tuning (Lester et al., 2021) and prefix tuning (Li & Liang, 2021), attempt to tackle this challenge by employing trainable prompts within a continuous space, resulting in impressive performance in transfer learning tasks. Prompts encapsulate task-specific knowledge with significantly fewer additional parameters compared to competing methods like Adapter (Pfeiffer et al., 2020; Wang et al., 2021) and LoRA (Hu et al., 2022).

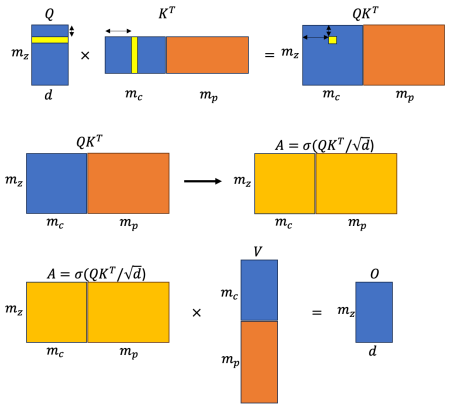


Figure 7: Illustration of Cross Attention with additional prompt.

Table 2: Cross-Attention Mechanisms

	Original		Concatenative		Addititive	
	Operation	Dim	Operation	Dim	Operation	Dim
Q	$W_q Z$	$b \times m_z \times d$	$W_q Z$	$b \times m_z \times d$	$W_q Z$	$b \times m_z \times d$
K	$W_k C$	$b \times m_c \times d$	W_k cat$(C, \text{repeat}(p, b))$	$b \times (m_c + m_p) \times d$	W_k $(C + \text{repeat}(p, b))$	$b \times m_c \times d$
V	$W_v C$	$b \times m_c \times d$	W_v cat$(C, \text{repeat}(p, b))$	$b \times (m_c + m_p) \times d$	W_v $(C + \text{repeat}(p, b))$	$b \times m_c \times d$
A	$\sigma(QK^T/\sqrt{d})$	$b \times m_z \times m_c$	$\sigma(QK^T/\sqrt{d})$	$b \times m_z \times (m_c + m_p)$	$\sigma(QK^T/\sqrt{d})$	$b \times m_z \times m_c$
O	AV	$b \times m_z \times d$	AV	$b \times m_z \times d$	AV	$b \times m_z \times d$

C Further Details on the Proposed Method

In this section, we will delve into discussing two different mechanisms for injecting the prompt into the cross-attention module, known as the concatenative mechanism and the additive mechanism. Their basic operations can be found in Table 2. In text-to-image diffusion models, the cross-attention layers are positioned to integrate textual embeddings into visual generation to regulate the output's concept. Hence, these specific layers are as the most suitable for prompt injection, aligning with our goal.

Concatenative Mechanism. In this mechanism, the prompt is concatenated with the textual embedding $\mathbf{C}$ before being used as the key and value matrix inputs to the cross-attention module. Let $\mathbf{p} \in \mathbb{R}^{[1, m_p, d_p]}$ be the additional learnable prompt, where $m_p = km_c$ is the prompt size and $d_p = d_c$ is the dimension of the prompt. The main difference compared to the original mechanism is the projected matrices K and V as shown in Table 2.

Softmax normalization is applied on the last dimension of the attention score matrix $\mathbf{A}$. In addition, the scaling factor $\sqrt{d}$ is used to prevent the attention score from being too small when the dimension of the latent vector is large (Vaswani et al., 2017).

One of the advantages of this mechanism is that adding the prompt does not change the mechanism of the cross-attention module, which means that there is no need for a new architecture and it does not interfere with the model's ability to generate good content. We can use the same pre-trained model and only need to modify the corresponding cross-attention module in the codebase. With $\mathbf{p} = c$, the model can generate the same output as the original model.

Secondly, this mechanism allows us to utilize a larger prompt (theoretically, the prompt can be arbitrarily large). By using a larger prompt, we can either remove more undesirable content or preserve more desirable content. Experiments in Section D.1.2 demonstrate that the performance of this mechanism is scalable with the size of the prompt.

Limitation of the concatenative mechanism: The main limitation of the concatenative mechanism is that it relies on softmax normalization to distribute the attribution from the additional prompt to the entire textual path. This issue is also illustrated in Figure 7. Because of the linearity of the projection operation, the matrices $\mathbf{K}$ and $\mathbf{V}$ can be decomposed into two parts separately, the

projections of original textual embedding $\mathbf{C}$ (blue color) and prompt $\mathbf{P}$ (orange color), i.e., $\mathbf{K} = [\mathbf{W}_k\mathbf{C}, \mathbf{W}_k\mathbf{P}], \mathbf{V} = [\mathbf{W}_v\mathbf{C}, \mathbf{W}_v\mathbf{P}]$. As a result, the dot-product between the query $\mathbf{Q}$ and the key $\mathbf{K}$ can also be decomposed into two disjointed parts, i.e., $\mathbf{Q}\mathbf{K}^T = [\mathbf{Q}(\mathbf{W}_k\mathbf{C})^T, \mathbf{Q}(\mathbf{W}_k\mathbf{P})^T]$. Without the softmax normalization, the output of the cross-attention module will be just the sum of the two disjointed parts, which is not desirable. The softmax normalization applied on the last dimension of the dot-product score matrix $\mathbf{Q}\mathbf{K}^T$ helps to distribute the attribution from the prompt to the entire textual path, enabling the model to attend to the additional prompt better. However, as the prompt size increases, the softmax normalization will distribute the attribution from the prompt to the entire textual path more evenly, which can lead to the model's inability to attend to the prompt effectively, as shown in Section D.1.2.

Additive Mechanism. This mechanism injects the additional prompt by directly adding it to the textual embedding $\mathbf{C}$ before being used as the key and value matrix inputs to the cross-attention module. This retains the same advantages as the concatenative mechanism, i.e., it does not change the mechanism of the cross-attention module. It also permits a deeper integration of the prompt into the textual path, which allows the model to attend to the prompt more effectively. However, a limitation of this mechanism is that it is not scalable since its size is fixed to the size of the textual embedding. We compare the performance of two mechanisms in Section D.1.1.

Alternative Prompting Mechanism. Beyond the two aforementioned mechanisms, we acknowledge that there are several potential prompting mechanisms that can be used to modify the cross-attention module. For example, we can inject the prompt before the text encoder, by using a learnable word embedding vector associated with a special token S^* to represent the prompt as in Textual Inversion (Gal et al., 2022). We can also amortize the prompt by using a learnable function to generate the prompt from the textual embedding, i.e., $\mathbf{p} = f(\mathbf{c})$. We leave the exploration of these mechanisms for future work.

D Further Experiments

D.1 Ablation Study

D.1.1 Concatenative vs Additive Mechanism

In this experiment, we conducted a comparison of erasing performance between two mechanisms. The evaluation was performed on a subset of 5 classes from the Imagenette dataset, including 'Cassette Player', 'Church', 'Garbage Truck', 'Parachute', and 'French Horn'. Additionally, we assessed the erasing performance in a nudity concept setting using the I2P dataset with the NER at a threshold of 0.5 as the erasure metric.

The results, presented in Table 3, indicate that the concatenative prompting mechanism outperforms the additive prompting mechanism in terms of erasing performance. This is evident in the 2.44% increase in ESR-1 and 3.16% increase in ESR-5, as well as a 0.3% decrease in NER. However, it is worth noting that the concatenative prompting mechanism is less effective in preserving unrelated concepts, as indicated by a drop of 2.8% in PSR-1 compared to the additive prompting mechanism.

While additive prompting theoretically provides a deeper integration between the prompt and the real textual input, the concatenative prompting mechanism has demonstrated greater effectiveness in erasing the target concept. Furthermore, its scalability allows for varying prompt sizes, a feature discussed in the next section. As a result of its superior erasing performance, we adopt the concatenative prompting mechanism as the default setting in all other experiments.

Table 3: Analytical results to different prompting mechanisms and prompt size.

Method	ESR-1↑	ESR-5↑	PSR-1↑	PSR-5↑	NER↓
Additive	96.40	92.32	84.48	97.92	1.7
Concat	98.84	95.48	81.68	97.56	2.0
k=1	98.60	96.04	84.76	97.56	2.17
k=10	98.84	95.48	81.68	97.56	1.70
k=100	99.68	97.08	82.68	96.84	**1.15**
k=200	99.60	96.80	77.24	94.16	1.49

D.1.2 EFFECT OF PROMPT SIZE

In this experiment, we explore the influence of prompt size on erasing performance by systematically varying the parameter k within the range of 1 to 200. The experimental setup mirros the previous experiment, where we focus on erasing object-related concepts with 5 classes from the Imagenette dataset, and also erasing nudity concepts with the I2P dataset.

It can be seen from the results in Table 3 that the erasing performance increases as the prompt size becomes larger, but becomes saturated once the prompt size becomes sufficiently large. Specifically, ESR-1 improves from 98.60% to 99.68%, and ESR-5 from 96.04% to 97.08% as the prompt size increases from 1 to 100. Similarly, the NER score decreases from 2.2 to 1.1 within the same range, indicating a consistent impact of prompt size across different types of concepts. However, the erasing performance is accompanied by a trade-off in preserving performance, as PSR-1 decreases from 84.76% to 77.24% when the prompt size increases from 1 to 200. The observed saturation in erasing performance for larger prompt sizes can be attributed to the softmax function in the cross-attention module, which becomes increasingly uniform and small as the prompt size grows. This makes it more challenging for the model to distinctly focus on the specific concept for erasure.

D.1.3 INFLUENCE OF HYPER-PARAMETER

In this experiment, we investigate the influence of the hyper-parameter λ on erasing performance of our method. We conduct the experiment to erase object-related concepts, using the same experimental setup as described in Section 4.6 of the main paper. Specifically, we focus on a subset of 5 classes from the Imagenette dataset, including 'Cassette Player', 'Church', 'Garbage Truck', 'Parachute', and 'French Horn' as the concepts to be erased, while preserving the remaining classes. We vary the hyper-parameter λ from 0.01 to 10.0. It is worth noting that the hyper-parameter λ is utilized to introduce the prompt in the knowledge removal stage, as described in Section 3.2 of the main paper. Therefore, it must be strictly positive, i.e., $\lambda > 0$.

The results depicted in Figure 8 reveal a clear decreasing trend in erasing performance as the hyper-parameter λ increases, while the preserving performance exhibits a slight increase. Specifically, the erasing performance peaks at $\lambda = 0.01$, with an ESR-5 of 96.8%, and drops significantly to 80.6% at $\lambda = 1.0$, and to around 40% at $\lambda = 10.0$. Conversely, the preserving performance, measured by PSR-5, increases from 97.3% to 98.0% as λ increases. The PSR-1 also exhibits a similar trend, increasing from 81.6% to around 86% as λ varies from 0.01 to 10.0.

This result aligns with our analysis in Section 3.2 of the main paper, where the hyper-parameter λ is employed to control the trade-off between erasing and preserving performance. In the knowledge removal stage, the L_2 term serves as a regularization term to minimize the change in the model's parameters, thereby preserving the knowledge encoded in the prompt of the erased concepts learned from the knowledge transfer stage. Therefore, a larger λ encourages the model to preserve the knowledge in the prompt more strongly, leading to smaller changes in the model's parameters and better preserving performance, but worse erasing performance. In other experiments, we use $\lambda = 0.1$ as the default value for the hyper-parameter λ.

D.1.4 WHERE TO INJECT PROMPT

In this paper, we have introduced two different prompting mechanisms: concatenative and additive prompting. While in the additive prompting, we add the prompt $\mathbf{i}$ directly to the textual embedding c as in Table 2, which can be understood as injecting the prompt in entire cross-attention layers of the U-Net, in the concatenative prompting, we need to specify where to inject the prompt.

In this experiment, we explore the influence of prompt injection at different layers of the model on erasing performance within the U-Net architecture of the Stable Diffusion model. The U-Net comprises three main components: the down-sample blocks, the middle block, and the up-sample blocks. Each of these components includes multiple cross-attention layers that can be used to inject the textual/conditional input, except for the middle block, which contains only one cross-attention layer.

We compare three different settings: injecting the prompt at the middle block, the middle and up-sample blocks, and the down-middle-up sample blocks (i.e., all cross-attention layers in the U-Net), with the same experimental setup as in previous experiments. The results in Table 4 indicate that

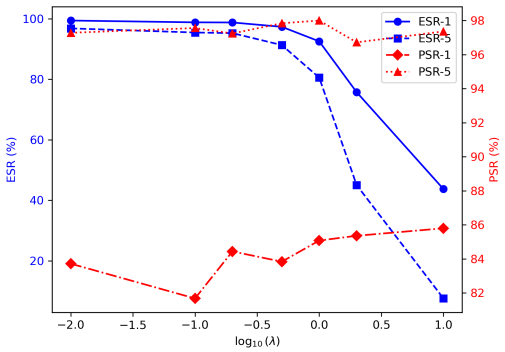


Figure 8: Impact of the hyper-parameter λ on the erasing performance.

Table 4: Where to inject the prompt.

layer	ESR-1↑	ESR-5↑	PSR-1↑	PSR-5↑
mid	98.84	95.48	81.68	97.56
mid-up	99.64	97.92	75.68	95.12
down-mid-up	99.65	98.36	59.04	86.28

injecting the prompt at all cross-attention layers in the U-Net yields the best erasing performance, albeit with a significant drop in preserving performance.

It is noteworthy that as the number of cross-attention layers used for prompt injection increases, erasing performance improves at the expense of preserving performance. The optimal trade-off between erasing and preserving performance is achieved by injecting the prompt at the middle block only. This setting was consequently chosen as the default for all subsequent experiments. It strikes a balance, demonstrating effective erasure while still preserving relevant elements in the input.

D.1.5 Further Results on Erasing Artistic Style Concepts

How to systematically evaluate the erasure performance? In this paper, we have conducted three sets of experiments to assess the performance of our proposed method against other erasure baselines. In the initial two sets, targeting the erasure of object-related and nudity concepts, we employed pre-trained detectors like ResNet-50 (He et al., 2016) and Nudenet (Praneet, 2019) to identify the presence of these concepts in the generated images. This systematic approach enabled us to evaluate the erasure performance rigorously. However, in the final experiment set, aimed at erasing artistic style concepts, we encountered a challenge: the absence of a pre-trained detector capable of accurately assessing the presence of an artistic style in the generated images. To address this challenge, previous studies (Gandikota et al., 2023) proposed human evaluations, which are subjective and time-consuming.

In this experiment, we explored the use of the CLIP alignment score as an alternative metric to evaluate erasure performance. Initially, we generated 1200 images from the original SD model using lengthy, specifically designed prompts (credited to (Gandikota et al., 2023)) to capture images with the artistic style of a particular artist. Subsequently, we measured the CLIP alignment score

Table 5: CLIP alignment score measured on the original SD model.

	Content & Artist	Artist	Content
Kelly McKernan	31.47 ± 2.58	27.67 ± 2.73	29.69 ± 2.43
Tyler Edlin	30.63 ± 2.22	23.67 ± 1.24	30.12 ± 2.49
Kilian Eng	29.87 ± 2.64	25.08 ± 1.31	30.54 ± 2.36
Thomas Kinkade*	34.63 ± 1.96	31.13 ± 2.38	31.09 ± 2.22
Ajin: Demi Human*	30.70 ± 2.55	27.65 ± 3.24	25.38 ± 2.77
VanGogh*	33.66 ± 2.41	30.36 ± 1.17	28.62 ± 3.28

between the generated images and three different textual inputs: the full prompt containing both the content and the artist name, the artist name alone, and the content alone. The results presented in Table 5 revealed intriguing insights. On one hand, when measuring based solely on the artist name, the CLIP alignment score was consistently the lowest in all cases, except for Thomas Kinkade, Ajin: Demi Human, and VanGogh. Conversely, when measuring based solely on the content, the CLIP alignment score was relatively higher. Lastly, when considering the full prompt, inclusive of both the content and the artist name, the CLIP alignment score was consistently the highest in all cases, except for Kilian Eng. This suggests that, from the CLIP's perspective, the generated images may not align well with just the artist name, but they exhibit strong alignment with the full prompt, encompassing both the content and the artist name. Consequently, to evaluate erasure performance, we can leverage CLIP as a zero-shot classifier, as highlighted in Section 3.4.

Qualitative Results. In addition to the quantitative results reported in Section 3.4, we provide further qualitative results as shown in series of figures from Figure 9 to Figure 14 to illustrate the erasure performance of our method and the baselines. Because of our internal policy on publishing sensitive content like nudity, we are able to provide results for the erasure of artistic style concepts and object-related concepts only.

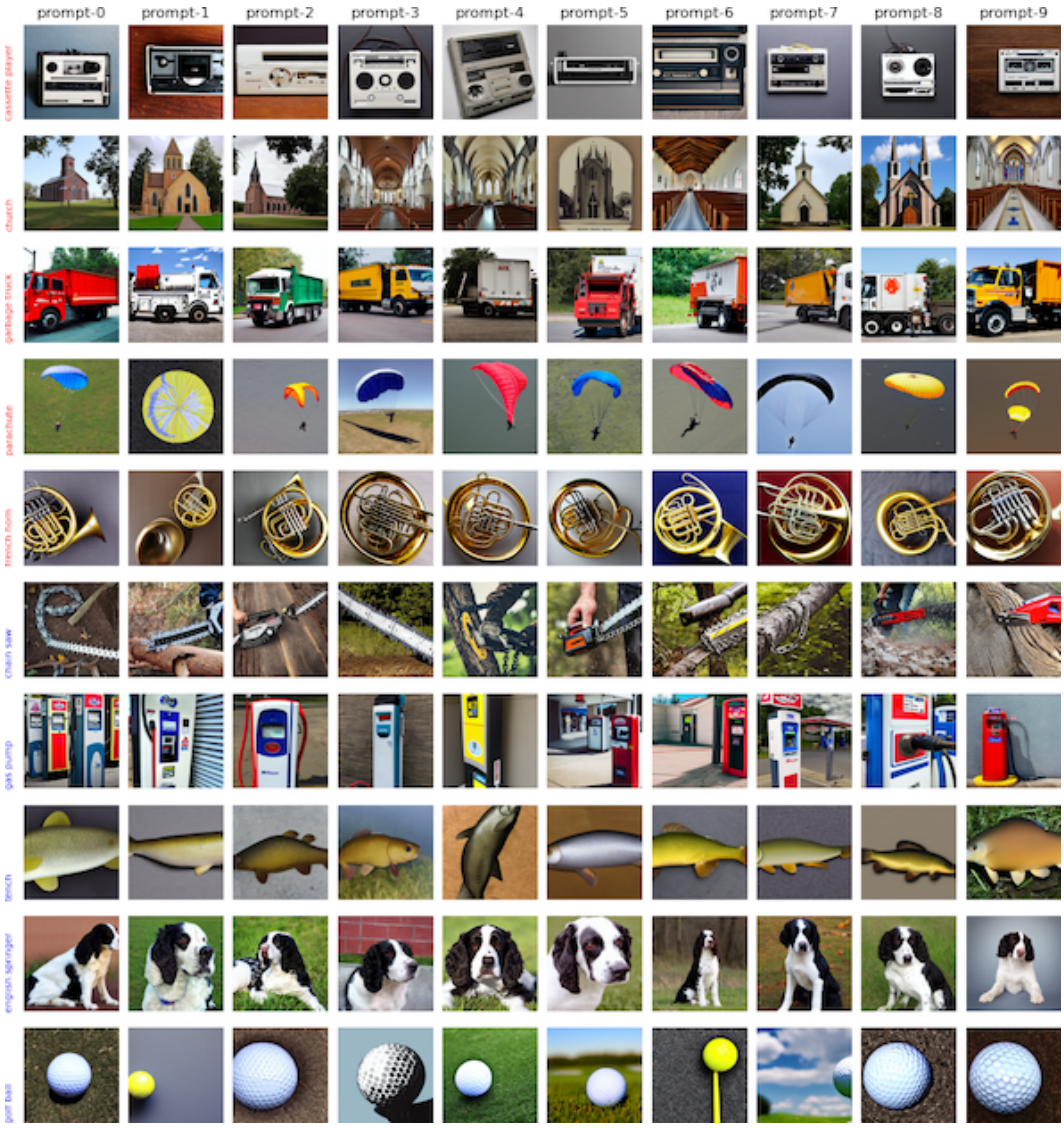


Figure 9: Generated images from the original model. Five first rows are to-be-erased objects (marked by red text) and the rest are to-be-preserved objects. Each column represents different random seeds.

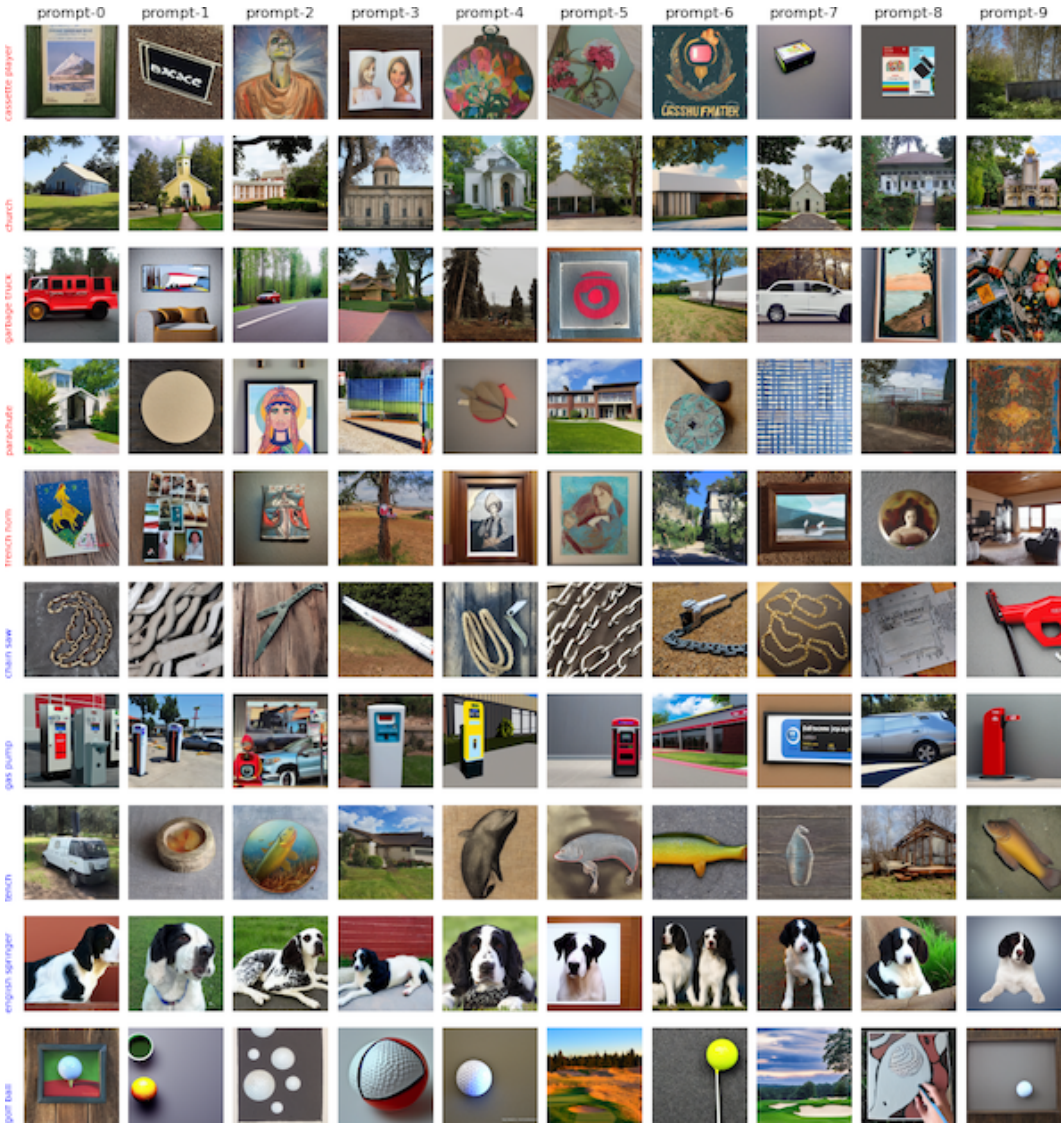


Figure 10: Erasing objects using ESD. Five first rows are to-be-erased objects (marked by red text) and the rest are to-be-preserved objects. Each column represents different random seeds.

Figure 11: Erasing objects using UCE. Five first rows are to-be-erased objects (marked by red text) and the rest are to-be-preserved objects. Each column represents different random seeds.

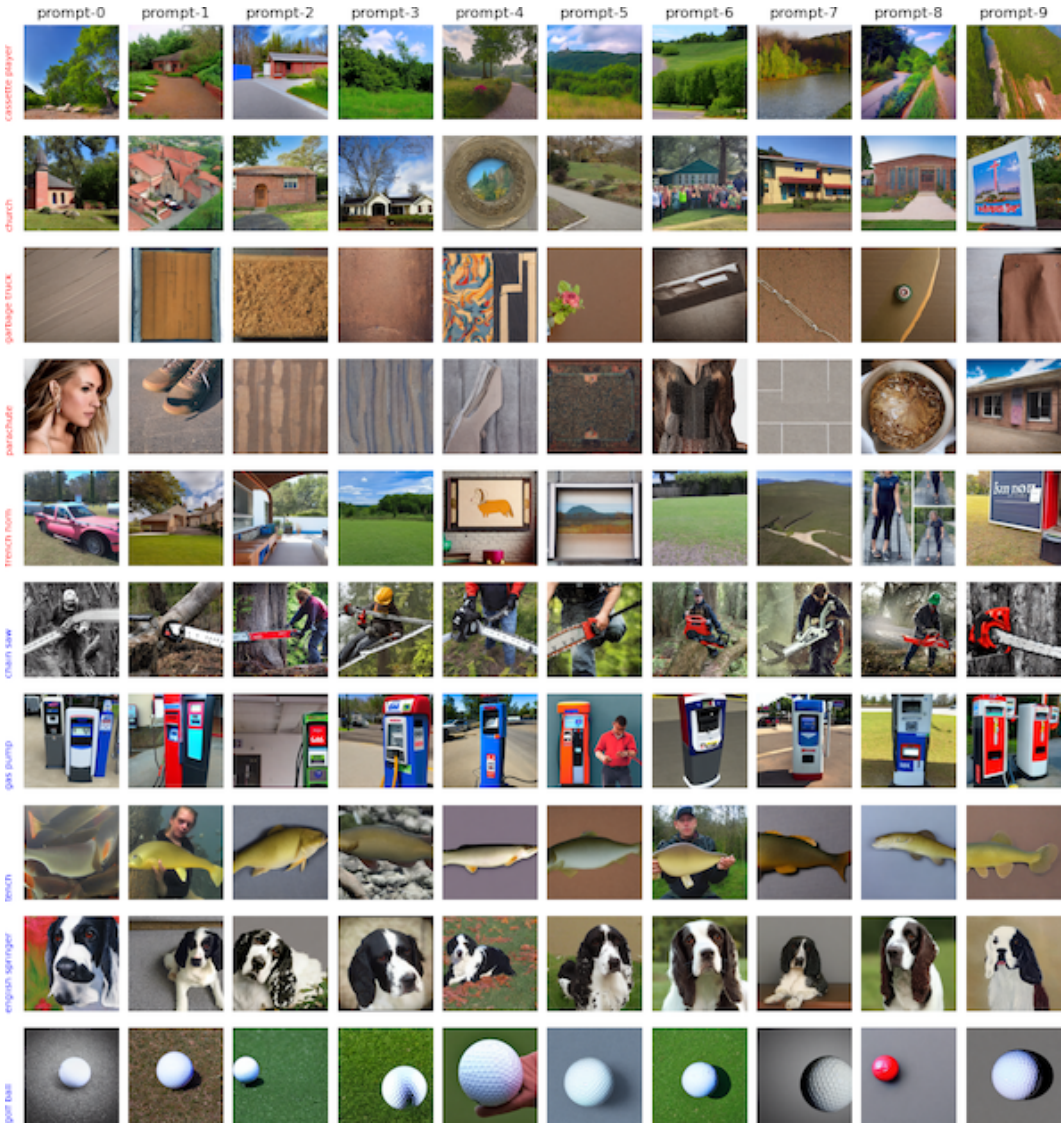
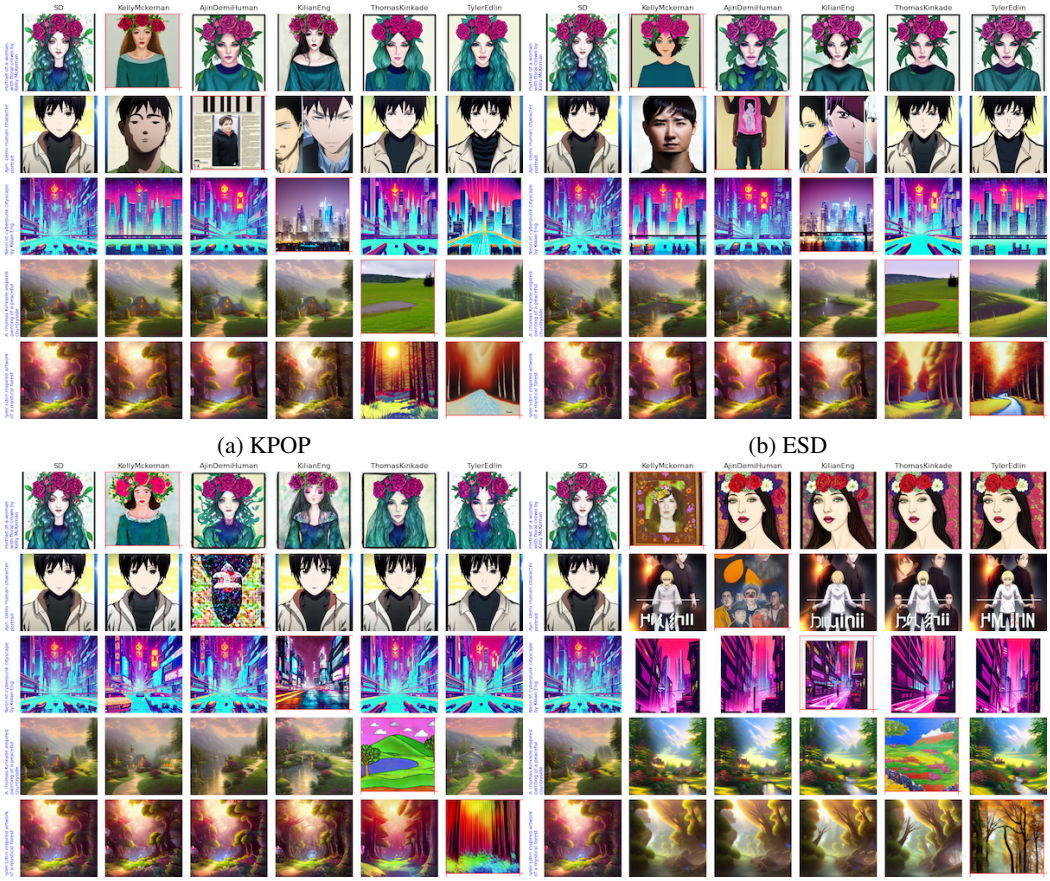


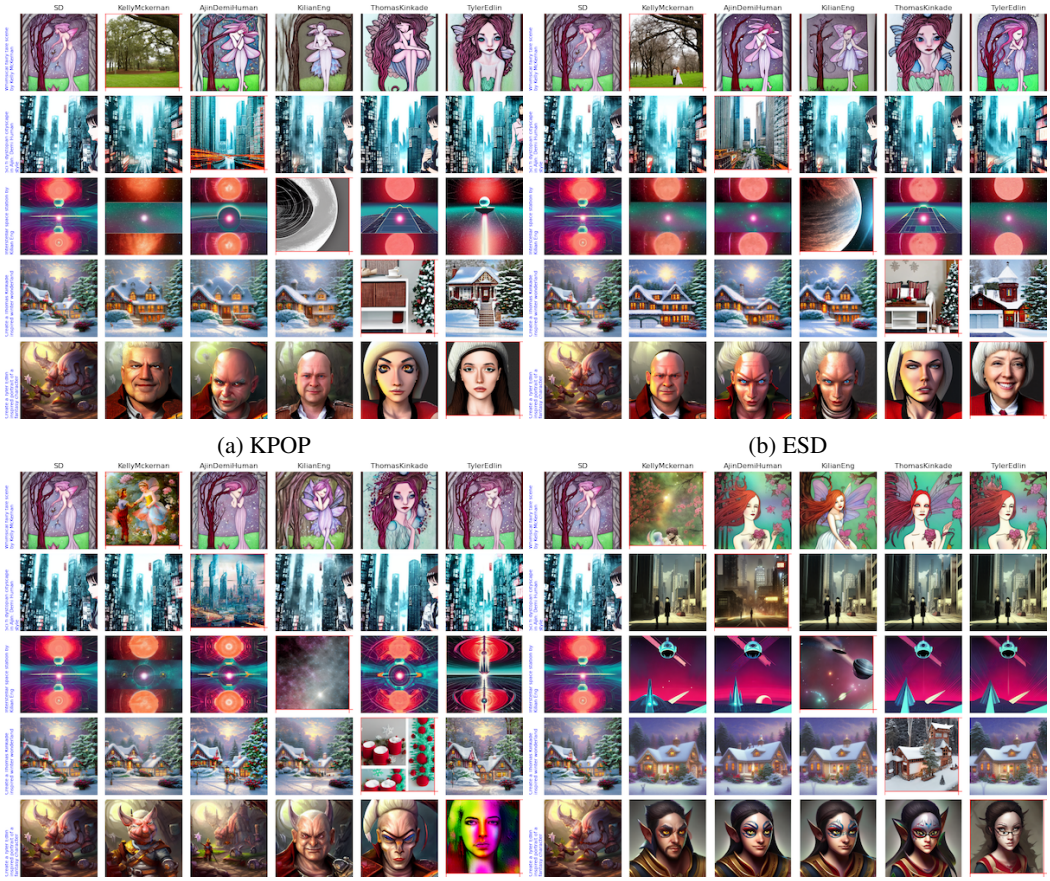
Figure 12: Erasing objects using our method (KPOP). Five first rows are to-be-erased objects (marked by red text) and the rest are to-be-preserved objects. Each column represents different random seeds.



(a) KPOP (b) ESD

(c) UCE (d) CA

Figure 13: Erasing artistic style concepts. Each column represents the erasure of a specific artist, except the first column which represents the generated images from the original SD model. Each row represents the generated images from the same prompt but with different artists. The ideal erasure should result in the change in the diagonal pictures (marked by a red box) compared to the first column, while the off-diagonal pictures should remain the same.



(a) KPOP (b) ESD

(c) UCE (d) CA

Figure 14: Erasing artistic style concepts (continue). Each column represents the erasure of a specific artist, except the first column which represents the generated images from the original SD model. Each row represents the generated images from the same prompt but with different artists. The ideal erasure should result in a change in the diagonal pictures (marked by a red box) compared to the first column, while the off-diagonal pictures should remain the same.