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

MapRoute: Semantic Routing for Precise Concept Erasure with Mapper

Sihao Li Baixi Liang Shuohong Xia Yunyun Yang*
Harbin Institute of Technology, Shenzhen
Shenzhen, China
*yangyunyun@hit.edu.cn

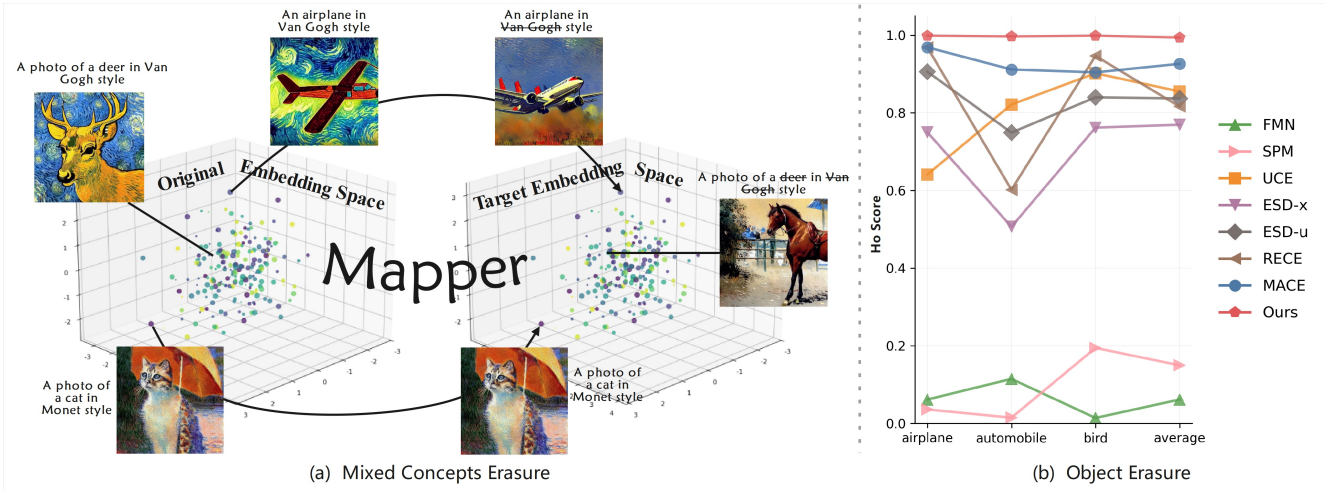

Figure 1. **Our method:** MapRoute, achieves precise concept erasure by learning a mapping from the original embedding space to the target embedding space. As illustrated in (a), MapRoute transforms embeddings from the original space (left) to the target space (right), demonstrating the successful erasure of the concepts of "deer" and "Van Gogh". In (b), we present quantitative results for object erasure across multiple categories, including "airplane", "automobile", "bird", and all 10 categories collectively. MapRoute consistently outperforms prior methods, with H_o representing the overall erasure capability (see Section 4.1 for details).

Abstract

Contemporary commercial and open-source diffusion models have demonstrated remarkable performance in text-to-image generation, enabling widespread applications in creative design and content creation. However, legitimate requirements often necessitate the removal of specific semantic concepts from pretrained models, such as copyright protection, privacy compliance, or personalized customization. Existing concept erasure methods suffer from two critical limitations: (1) Incomplete suppression, where the model still occasionally generates images containing the target concept; (2) Poor semantic selectivity, which degrades the generation quality of unrelated concepts and compromises overall model utility. To address these challenges, we propose `MapRoute`*, a lightweight, semantics-aware concept erasure framework based on dynamic routing. Our approach introduces a set of modular components placed after a frozen pretrained text encoder, termed Mappers. Each Mapper learns a linear mapping from a target concept to a surrogate concept. During inference, the system dynamically activates the top-K Mappers most relevant to the input prompt, based on cosine similarity between the text embedding and all the target concept embeddings, and applies their transformations sequentially. This input based, modular intervention enables precise, input based erasure while avoiding unnecessary interference with irrelevant semantics. Extensive experiments demonstrate that* `MapRoute` *effectively suppresses specified concepts while significantly reducing collateral damage to unrelated concept. Moreover,* `MapRoute` *outperforms state-of-the-art baselines in terms of generation fidelity, semantic consistency, and scalability to multi-concept erasure scenarios.Code is available at https://github.com/GG-li/MapRoute.*

1. Introduction

Large-scale text-to-image (T2I) models [5, 6, 13, 25, 28, 29] have demonstrated remarkable capabilities in rapidly generating high-quality and contextually coherent images. However, this extensive generative power has raised significant concerns regarding potential risks such as copyright infringement, privacy violations, and the dissemination of

Figure 2. **Comparison of generated results:** We select four Mapper modules trained with distinct surrogate concepts and generate images using the identical prompt "Cameron Diaz in an official photo" to evaluate the model's insensitivity to the choice of surrogate concept.

prohibited content [22, 26, 30, 33]. To mitigate these ethical and legal hazards, it is necessary to prevent models from generating specific sensitive or undesirable concepts—a process commonly referred to as *concept erasure* [1, 23, 27].

Existing concept erasure methods predominantly rely on full or partial fine-tuning of the model's original parameters [11, 15]. While effective to some extent, these approaches frequently induce unintended distortions or artifacts in non-target concepts, degrading the overall generation quality.

Recent advancements can be broadly categorized into three paradigms: *Loss-Based Optimization*, *Closed-Form Projections*, and *Plug-in Adapters* [38]. Given the evolving nature of societal norms and regulatory requirements, adaptive mitigation strategies are essential. Among these, *Plug-in Adapters* strike a promising balance between customizability and transferability: when new sensitive concepts emerge, only the adapter needs fine-tuning or replacement, without retraining the entire foundation model. Nevertheless, existing Plug-in Adapter methods suffer from two key limitations.

One limitation is that they typically require concept-specific training and depend critically on high-quality paired data (e.g., 'target concept' vs. 'surrogate concept') to precisely localize semantic representations for suppression [9, 21, 32]. Without such data, adapters may fail, leading to either incomplete erasure or unintended inhibition of legitimate concepts. A further constraint is that when erasing multiple concepts simultaneously, parameter conflicts across different adapters can result in partial erasure failure or noticeable quality degradation, necessitating complex parameter isolation mechanisms[15, 17, 37, 42].

To address these challenges, we propose **MapRoute**, a novel framework that performs concept erasure by editing the text encoder's embedding vectors prior to image generation. Our approach features two key innovations:

First, we introduce a lightweight module, dubbed **Mapper**, which learns a conditional identity mapping via a two-stage training scheme. The Mapper learns a conditional identity mapping from the original text embedding space to a rectified space via a two-stage training scheme: (1) A self supervised stage where it learns an identity mapping to preserve all semantic content; (2) A concept-specific stage where it learns to map embeddings of target concepts to surrogate concepts. Through this two phase optimization, the module achieves almost zero distortion on the concepts in the two phases while effectively eliminating the designated ones. This design enables the simultaneous and precise erasure of multiple concepts, ensuring flexible adaptation to diverse scenarios. Moreover, since the Mapper is designed to learn the mapping of target concepts, it breaks free from the reliance on high-quality paired data. Specifically, choosing different surrogate concepts has no impact on the erasure effect of the target concept.As shown in Figure 2.

Second, to handle parameter conflicts and catastrophic forgetting[20] when erasing a large number of concepts, we incorporate an input-driven routing mechanism. This mechanism operates on a *top-k* principle: it calculates the cosine similarity between the input prompt embedding and all target concept embeddings, dynamically selecting the most relevant mappers for serial execution.

Both qualitative and quantitative evaluations demonstrate that MapRoute successfully erases a wide range of sensitive content, including specific objects, artistic styles, and celebrity likenesses. Compared to state-of-the-art (SOTA) methods, our approach achieves more thorough concept erasure and superior preservation of non-target semantics. These advantages are particularly prominent in the scenario of erasing mixed concepts. Extensive experiments consistently demonstrate that MapRoute outperforms existing SOTA methods in both single- and multi-concept erasure tasks, setting a new benchmark in semantic consistency, erasure completeness, and generation fidelity.

2. Related Works

In recent years, to achieve selective suppression of harmful or sensitive concepts in text-to-image (T2I) diffusion models, researchers have proposed various concept erasure methods based on different optimization strategies. Existing works can be systematically categorized into three mainstream paradigms: loss-function-based optimization, closed-form analytical projection, and plug-and-play adapter modules [38, 40, 43].

Loss-function-based optimization guides the model to weaken the generative capability of the target concept during fine-tuning. Representative works include ESD[7], which minimizes the discrepancy between conditional and unconditional noise predictions to achieve concept neu-

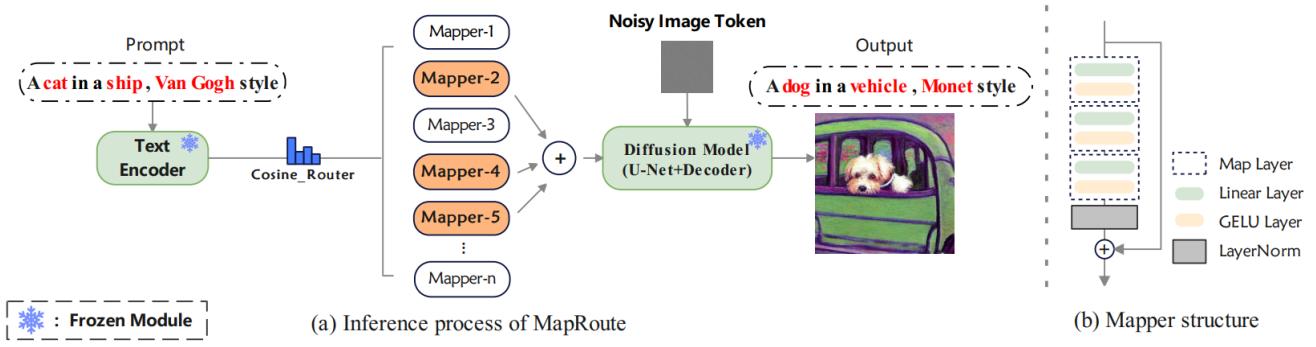

Figure 3. **Model Architecture:**(a) Inference pipeline of MapRoute, demonstrating the erasure of the concepts "cat", "ship", and "Van Gogh". (b) Architecture of the Mapper module, where each Map Layer consists of a linear layer followed by a GELU activation.

tralization; CRCE[39] and CORE[45], which further introduce anchor concepts and preservation set constraints to balance erasure efficacy and model generality; AGE[4] and FADE[34], which employ multi-objective optimization frameworks to suppress the target concept while explicitly preserving semantically adjacent or contextually relevant benign content. While offering high flexibility, these methods typically require extensive iterative training and involve non-negligible adjustments to model parameters, which may unpredictably impact the overall generative capability of the model.

Closed-form analytical projection methods operate under the assumption that the target concept occupies a specific subspace within the model's internal representations. These methods directly modify parameters via analytical computations, such as least squares or Singular Value Decomposition (SVD). UCE[8] pioneered this approach by performing closed-form updates on the K/V matrices in cross-attention layers for training-free concept substitution. Subsequent works, such as MACE[19], scaled this to enable erasure of hundreds of concepts simultaneously. RealEra[18] enhanced robustness through neighborhood concept mining, while CE-SDWV[35] constructed principal component subspaces within the text encoder's embedding space for orthogonal projection. Although computationally efficient and highly reproducible, these methods struggle to precisely isolate the target subspace when the target concept exhibits non-linear semantic entanglement with legitimate concepts. For instance, using SVD to extract and nullify the principal components associated with *English Springer Spaniel* might inadvertently suppress generic features of *canines*, significantly impairing the model's ability to generate other dog breeds.

Plug-and-play adapter modules represent a lightweight, reversible erasure paradigm that requires no modification of the original model weights. Typical solutions involve inserting Low-Rank Adaptation (LoRA) modules[11], gating units, or residual attention gates into the UNet or cross-attention layers. For example, SPM[20] introduces semi-permeable LoRA adapters into the UNet, achieving targeted suppression by updating only a minimal set of parameters. Receler[12] combines concept localization regularization with adversarial training to improve the robustness of local erasure. These methods offer flexible deployment and are well-suited for dynamic safety policy updates. However, they often require supervised signals (e.g., pairs of safe/unsafe images) for training. Moreover, when erasing multiple target concepts concurrently, parameters from different adapters can conflict, leading to partial erasure failure [44] or degradation in the quality of generated images.

3. Method

Our goal is to develop a framework that accurately removes target concepts from T2I generation models. As illustrated in Figure 1, our MapRoute framework consists of three components: (1) the Mapper module design, (2) the learning strategy, and (3) the semantic routing design. The framework takes two sets of inputs: a discretized representation of embedding vectors in the semantic space, and a set of target concept-surrogate concept pairs(the concept to be removed and a benign replacement). It outputs a fine-tuned model that dynamically selects the appropriate Mapper module to integrate, ensuring that the model no longer generates images corresponding to prompts containing the target concept, while preserving nearly no impact on images generated from prompts that do not contain the target concept.

Why Should We Insert a New Module Instead of Fine-tuning Existing Layers? Current mainstream methods for concept erasure involve fine-tuning linear or attention layers; although effective at eliminating target concepts, these approaches modify the model's parameters globally, degrading the generation quality for unrelated concepts.

3.1. The structural design of MapRoute

In T2I diffusion models, prompts are encoded by a text encoder and used as conditions to iteratively denoise random noise in the latent space, ultimately generating images. Mapper is inserted after the text encoder and learns a mapping from the original embedding space to another embedding space, thereby preserving the generation quality for unrelated concepts while effectively removing the target concept.

We introduce a lightweight feedforward module $M_{c_{tar}}$, featuring three Map layers and a LayerNorm module (see Figure 3), designed to enhance semantics with minimal parameters. In our T2I diffusion model, it is inserted post-text-encoder to learn a Conditional Identity Mapping that suppresses undesirable encodings passed to the U-Net, as formulated below:

$$M_{c_{tar}}(E(c)) = \begin{cases} E(c_{sur}), & c = c_{tar} \\ E(c), & c = others \end{cases} \tag{1}$$

The notation is defined as follows: C_{tar} refers to the set of target concepts, with each constituent concept designated as c_{tar}. A specialized mapper module $M_{c_{tar}}$ is assigned to each c_{tar} to achieve its eradication. Here, $E(\cdot)$ is a frozen, pre-trained text encoder. The variable c represents the embedding vector of a prompt encoded by $E(\cdot)$, and c_{sur} is the embedding for the surrogate concept.

This modular design, which tailors a Mapper for each specific concept, enables flexible customization for multi-concept scenarios and facilitates the targeted deployment of erasure strategies on pre-trained diffusion models. Moreover, it effectively prevents the catastrophic forgetting issue that arises from sequentially chaining a large number of Mappers.

3.2. Identity mapping learning

Inspired by the concept dictionary design in Splice[2], we leverage the insight that the embedding space of the CLIP[14] encoder exhibits a semantically linear structure. This implies that an image can be approximately represented as a sparse, non-negative linear combination of human-intelligible semantic concepts. Consequently, the semantic content that CLIP can express is approximately covered by a finite concept dictionary. To construct this vocabulary, we considered the most frequent one-word and two-word concepts in the text captions of the LAION-400M dataset[31]. After removing NSFW samples and pruning the set to ensure no two concept embeddings had a cosine similarity greater than 0.9, we ultimately selected the top 10,000 most frequent single-word concepts and the top 5,000 most frequent two-word concepts to form our dictionary, thereby achieving a discrete representation of the embedding space. To enable precise concept erasure, we devised a two-stage optimization strategy. The entire algorithm is summarized with pseudo-code in Algorithm 1, and a detailed explanation follows.

The first stage involves freezing the text encoder's parameters and training Mapper to learn an identity mapping for the embeddings of all concepts in the dictionary.

$$\mathcal{L}_{stage1} = \mathbb{E}_{c \in C}\left[\|M_{c_{tar}}(E(c)) - E(c)\|_2^2\right] \tag{2}$$

Let C be the concept dictionary and c is an element within it. Since the T2I model's text encoder produces embeddings of a fixed dimension, we assessed the mapper's performance by applying it to 1,000 random 768-dimensional vectors that simulate real text embeddings. After just 10 epochs of training, the Mapper achieved an average MSE of 1×10^{-6} between the input and output. This result confirms that the Mapper has effectively learned a precise identity mapping after the first stage. Detailed ablation experiments are presented in Section 4.5 and the Appendix.

3.3. Target-surrogate mapping learning

The objective of the first training stage is to learn an identity mapping for all possible prompts. Building upon this foundation, the second stage focuses on learning the mapping relationship between target and surrogate concepts. Target concept is the concept intended for erasure, while surrogate concept is a predefined concept used to replace it during the erasure process, forming a concept pair. Previous approaches suffer from significant limitations in selecting these surrogates. For instance, methods like SPM [20] require manually specifying a unique surrogate for each target concept, a process that often demands extensive testing. Although a common strategy is to select a "semantic neighbor" of the target, defining such neighbors is highly non-trivial for a vast number of concepts. For example, it is challenging to determine if one artist's style is similar to another's or to find a concept semantically close to a specific cartoon character.

Since Mapper learns a conditional identity mapping, the choice of a surrogate concept is not critical. For any given target concept, the mapper can be trained to map it to a surrogate that has no semantic relation whatsoever, such as mapping "automobile" to "Van Gogh" or "truck" to "cat". Crucially, the empirical results indicate that the selection of different surrogate concepts yields largely consistent erasure performance, with the final outcome being largely independent of this choice(see Appendix). We formalize this objective as follows:

$$\mathcal{L}_{learn} = \mathbb{E}_{c_{tar}}\left[\|M_{c_{tar}}(E(c_{tar})) - E(c_{sur})\|_2^2\right] \tag{3}$$

$$\mathcal{L}_{keep1} = \mathbb{E}_{c_j \sim C}\left[\|M_{c_{tar}}(E(c_j)) - E(c_j)\|_2^2\right] \tag{4}$$

To achieve more precise erasure for celebrity and artist concepts, we retrieved an initial set of 10,000 celebrity records

by querying the Wikidata SPARQL endpoint[36] for entities that are instances of "human" and possess an "occupation" attribute. We then applied a language filter to the labels of these records, retaining only those with English labels and removing all non-English names. After this filtering step, a final set of 8,578 valid English person names was obtained. Following the design methodology of $\mathcal{L}_{keep1}$, we subsequently designed $\mathcal{L}_{keep2}$ based on this list.

$$\mathcal{L}_{keep2} = \mathbb{E}_{n_j \sim W} \left[\| M_{c_{tar}}(E(n_j)) - E(n_j) \|_2^2 \right] \quad (5)$$

Let W denote the set of all English person names, the overall loss function for the second stage is then formulated by combining the three components using two balancing hyperparameter, α and β (all set to 1 by default), as follows:

$$\mathcal{L}_{stage2} = \mathcal{L}_{learn} + \alpha \times \mathcal{L}_{keep1} + \beta \times \mathcal{L}_{keep2} \quad (6)$$

Following the two-stage loss training, Mapper can accurately erase the target concepts and maximally preserve unrelated ones.

Algorithm 1 Training Method of Mapper Module $M_{c_{tar}}$

Input: Pre-trained frozen text encoder $E(\cdot)$, Mapper module $M_{c_{tar}}(\cdot)$ (to be trained), target concept set C_{tar}, surrogate concept set C_{sur}, concept dictionary C, name set W, learning rate η, weights α, β
Output: Trained Mapper module $M_{c_{tar}}$

for epoch = 1 to T **do**
  $M_{c_{tar}}$.train()
  if epoch < 10 **then**
    Stage 1: Preserve all concepts
    for each batch of concepts $c \in C$ **do**
      $\mathcal{L}_{\text{stage1}} \leftarrow \| M_{c_{tar}}(E(c)) - E(c) \|_2^2$
      $M_{c_{tar}} \leftarrow M_{c_{tar}} - \eta \cdot \nabla_{\theta_M} \mathcal{L}_{\text{stage1}}$
    end for
  else
    Stage 2: Map target concepts to surrogate
    for each batch of target concepts $t \in C_{\text{tar}}$ **do**
      Sample $c_{\text{sur}} \sim C_{\text{sur}}$
      $\mathcal{L}_{\text{learn}} \leftarrow \| M_{c_{tar}}(E(t)) - E(c_{\text{sur}}) \|_2^2$
      Sample two thousand c_j,n_j
      $\mathcal{L}_{\text{keep1}} \leftarrow \| M_{c_{tar}}(E(c_j)) - E(c_j) \|_2^2$
      $\mathcal{L}_{\text{keep2}} \leftarrow \| M_{c_{tar}}(E(n_j)) - E(n_j) \|_2^2$
      $\mathcal{L}_{\text{stage2}} \leftarrow \mathcal{L}_{\text{learn}} + \alpha \cdot \mathcal{L}_{\text{keep1}} + \beta \cdot \mathcal{L}_{\text{keep2}}$
      Compute $\nabla_{\theta_M} \mathcal{L}_{\text{stage2}}$
      $M_{c_{tar}} \leftarrow M_{c_{tar}} - \eta \cdot \nabla_{\theta_M} \mathcal{L}_{\text{stage2}}$
    end for
  end if
end for

3.4. Semantic Route

By specifying different target concepts, we have trained a large number of Mappers, which collectively form a comprehensive erasure corpus. To enable a single model to flexibly erase multiple concepts from a prompt—while avoiding the issue of catastrophic forgetting that arises from simply integrating all Mappers—and to prevent the need for repeating the whole erasing pipeline each time for a dedicated model, we designed a corresponding routing mechanism. This mechanism retrieves and connects the k most relevant Mapper modules to the model based on the input prompt.

$$\mathcal{M}^{(k)} = \{ M_i \in \mathcal{M} \mid sim_i \in TopK\left(\{sim_1, sim_2, \dots\}, k\right) \} \quad (7)$$

Let $\mathcal{M}$ denote the set of all mapper modules, where M_i represents the module responsible for erasing the i-th concept. Let sim_i signify the cosine similarity between the i-th concept in $\mathcal{M}$ and the input prompt. The function $TopK(S, k)$ returns the set of the k largest elements from S.

The k selected modules are connected in series following the text encoder. When a prompt semantically activates these k mappers, the encoded representation passes through each activated module sequentially, thereby eliminating the corresponding unwanted concepts one after another.

4. Experiments

We conducted extensive experiments to comprehensively evaluate our proposed method and compare it against state-of-the-art (SOTA) baselines. Specifically, our experimental evaluation includes object concept erasure (Section 4.1), celebrity concept erasure (Section 4.2), artistic style erasure (Section 4.3), and mixed concepts erasure (Section 4.4). We compared our method against six recent baselines designed for these tasks: MACE[19], FMN[41], UCE[8], SPM[20], ESD[7], RECE[12], and GLoCE[16]. Finally, we performed ablation study (Section 4.5) to understand the impact of key components. Image quality generation tests and additional experiments are provided in Appendix.

In our implementation, all experiments were conducted on Stable Diffusion v1.4[28]. On a single NVIDIA A100, training a Mapper takes 13.3 minutes. Inference time is 6 seconds. For each text prompt, we used the default PNDM scheduler for sampling with 50 inference steps and a classifier-free guidance scale of 7.5. To ensure reproducibility, a fixed random seed was used for all generation processes, and the safety checker was disabled.

4.1. Object Erasure

In this section, we evaluate the effectiveness of our method on the CIFAR-10 dataset[14] under the experimental protocol established by MACE. For each target class, we fine-tune the model to erase one object category and generate 200 images. We then classify these generated images using CLIP[24], where a lower classification accuracy (ACC_e) indicates a more successful erasure of the target concept.

Table 1. **Quantitative results on object erasure.** Following the experimental setup of MACE, we conducted object erasure tests on the ten classes of the CIFAR-10 dataset, and some of the corresponding experimental results.

Method	Airplane Erased				Automobile Erased				Bird Erased				Average across 10 Classes			
	$\text{Acc}_e \downarrow$	$\text{Acc}_s \uparrow$	$\text{Acc}_g \downarrow$	$H_o \uparrow$	$\text{Acc}_e \downarrow$	$\text{Acc}_s \uparrow$	$\text{Acc}_g \downarrow$	$H_o \uparrow$	$\text{Acc}_e \downarrow$	$\text{Acc}_s \uparrow$	$\text{Acc}_g \downarrow$	$H_o \uparrow$	$\text{Acc}_e \downarrow$	$\text{Acc}_s \uparrow$	$\text{Acc}_g \downarrow$	$H_o \uparrow$
FMN	96.76	98.32	94.15	6.13	95.08	96.86	79.45	11.44	99.46	98.13	96.75	1.38	96.96	96.73	82.56	6.13
SPM	97.50	99.00	97.67	3.58	99.50	99.17	81.00	1.45	91.00	**99.61**	70.00	19.42	95.00	**99.53**	83.36	14.95
UCE	40.32	98.79	49.83	64.09	4.73	99.02	37.25	82.12	10.71	98.35	15.97	90.18	13.54	98.45	23.18	85.48
ESD-x	91.11	97.15	32.28	74.98	59.68	98.39	58.83	50.62	18.57	97.24	40.55	76.17	26.93	97.32	31.61	76.91
ESD-u	7.38	85.48	5.92	90.57	30.29	91.02	32.12	74.88	13.17	86.17	20.65	83.98	18.27	86.76	16.26	83.69
RECE	1.00	98.06	6.33	96.85	47.00	**99.83**	52.33	60.16	6.00	95.94	6.00	94.64	21.55	98.32	22.92	81.59
MACE	9.06	95.39	10.03	96.85	6.97	95.18	14.22	91.15	9.88	97.45	15.48	90.39	8.49	97.35	10.53	92.61
Ours	**0.00**	**99.56**	**0.00**	**99.85**	**0.50**	99.78	**0.33**	**99.65**	**0.00**	99.56	**0.00**	**99.85**	**0.50**	**99.53**	**0.92**	**99.37**
SD v1.4	96.06	98.92	95.08	–	95.75	98.95	75.91	–	99.72	98.51	95.45	–	98.63	98.63	83.64	–

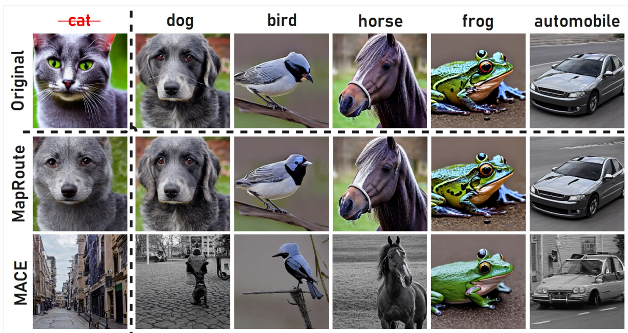

Figure 4. Samples of some classes after erasing the concept of "cat" using a fixed random seed of 447557. MACE, the current SOTA method, exhibits noticeable degradation on unrelated concepts. MapRoute demonstrates superior preservation performance, maintaining high visual fidelity across non-target classes.

To assess the generalization capability of concept erasure, we further follow MACE and construct three semantic synonyms and a surrogate concept for each target class (see Appendix for the full list). For each synonym, we generate 200 images and measure the resulting CLIP classification accuracy (ACC_g). A lower ACC_g reflects stronger generalization.

To evaluate preservation of non-target concepts, we generate 200 images for each of the remaining nine classes and compute their CLIP classification accuracy (ACC_s). A higher ACC_s indicates better preservation of the original generation quality for unrelated concepts, confirming minimal unintended side effects.

To thoroughly evaluate erasure, we expand each CIFAR-10 class into a set of ten semantic sub-concepts (e.g., for 'cat', we include 'kitten', 'siamese', etc., generated by GPT-3.5[3]), treating them collectively as the target set. The full list is in the Appendix.

To evaluate the overall concept erasure capability of our method, we adopt the harmonic mean of ACC_e, ACC_g, and ACC_s as the primary metric, which is computed as follows:

Table 2. **Comparison of baselines and the proposed method** for image fidelity on text prompts containing target celebrities and remaining celebrities.

Method	Adam Driver			Adriana Lima			50 Celebrities		
	$\text{GCD}_e \downarrow$	$\text{GCD}_s \uparrow$	$H_c \uparrow$	$\text{GCD}_e \downarrow$	$\text{GCD}_s \uparrow$	$H_c \uparrow$	$\text{GCD}_e \downarrow$	$\text{GCD}_s \uparrow$	$H_c \uparrow$
FMN	21.20	51.52	62.30	22.00	64.24	70.45	–	–	–
SPM	2.80	81.36	88.58	3.20	80.95	88.17	–	–	–
GLoCE	**0.00**	79.40	88.52	81.60	91.16	30.62	3.20	79.28	87.17
MACE	0.80	88.56	93.58	**0.00**	89.20	94.29	9.50	81.83	85.95
Ours	**0.00**	**91.16**	**95.38**	**0.00**	**91.44**	**95.53**	**0.00**	**90.16**	**94.83**
SD v1.4	93.20	91.12	–	83.60	91.12	–	92.64	91.22	–

$$H_o = \frac{3}{(1-\text{ACC}_e)^{-1} + (\text{ACC}_s)^{-1} + (1-\text{ACC}_g)^{-1}} \quad (8)$$

An ideal method should minimize ACC_e and ACC_g while maximizing ACC_s and H_0. As shown in Table 1, FMN and SPM fail to fully erase the target class, evidenced by their relatively high ACC_e. In contrast, ESD-u and ESD-x over-regularize the model, slightly degrading ACC_s and indicating unintended interference with non-target classes. Although RECE and MACE achieve competitive performance, they still fall short of complete erasure, as reflected by non-negligible ACC_e and ACC_s values. Our method, by contrast, demonstrates superior precision, controllability, and generalization in object erasure, establishing it as a better approach for concept erasure in diffusion models.

4.2. Celebrity Erasure

To evaluate the effectiveness of MapRoute in erasing celebrity-related concepts, we selected several celebrity concepts from the list of 200 celebrities provided by MACE for single-concept erasure, and an additional 50 celebrities for multi-concept erasure. For single-concept erasure, following the setup in MACE, we used five templates to generate 250 images for each target celebrity concept and 2,500 images for 100 unrelated celebrity concepts. We evaluate the effectiveness of concept erasure (GCD_e) and the preservation of irrelevant celebrity concepts (GCD_s) using the GIPHY Celebrity Detector (GCD)[10], which achieves remarkable recognition accuracy. Furthermore, to compre-

hensively evaluate the performance of the model, we employ H_c as follows:

$$H_c = \frac{2}{(1 - GCD_e)^{-1} + (GCD_s)^{-1}} \tag{9}$$

However, unlike MACE, our evaluation metric differs in order to more thoroughly assess the extent to which celebrity concepts have been erased. The criterion for successful erasure is defined as a significant reduction in the top-1 accuracy of GCD in correctly identifying the target celebrity that has been erased.Detailed generation settings and other results are provided in the Appendix.

As shown in Table 2, which presents quantitative results for both single- and multi-concept erasure, MACE and SPM demonstrate satisfactory performance in removing target concepts, but with slight concept erosion. GLoCE fails to effectively erase certain concepts (e.g., Adrian Lima), while FMN not only shows inadequate erasure but also introduces noticeable concept erosion. These observations indicate that most existing models struggle to balance the trade-off between "erasure" and "preservation." In contrast, our MapRoute achieves GCD_e values close to 0 across different celebrity concepts, with GCD_s scores nearly identical to those of the original model[28]. This suggests that MapRoute successfully eliminates target concepts entirely while effectively preserving the integrity of other concepts.

4.3. Artistic Style Erasure

To evaluate the effectiveness of MapRoute in erasing artistic styles, we conducted experiments following a similar protocol to the celebrity concept erasure tests. From the list of 200 artists provided by MACE, we selected several artists for single-style erasure testing and another 50 artists for multi-style erasure testing.

For single-style erasure, adhering to the setup in MACE, A total of 250 images were generated for each target artistic style and 2,500 images for 100 unrelated artistic styles.

Consistent with MACE, two metrics were employed: $CLIP_e$ and $CLIP_s$. $CLIP_e$ measures the CLIP score between the erased artist's style and the images generated from the corresponding style prompts—a lower score indicates better erasure performance. $CLIP_s$ evaluates the alignment between the preserved artistic styles and the images generated from their corresponding prompts—a higher score suggests less impact on unrelated concepts, reflecting better preservation.

To further assess the overall erasure capability of the method, we used the difference of $CLIP_e$ and $CLIP_s$, calculated as follows:

$$H_a = \text{CLIP}_s - \text{CLIP}_e. \tag{10}$$

Detailed experimental settings and the results of multi-style erasure can be found in the Appendix. Table 3 presents

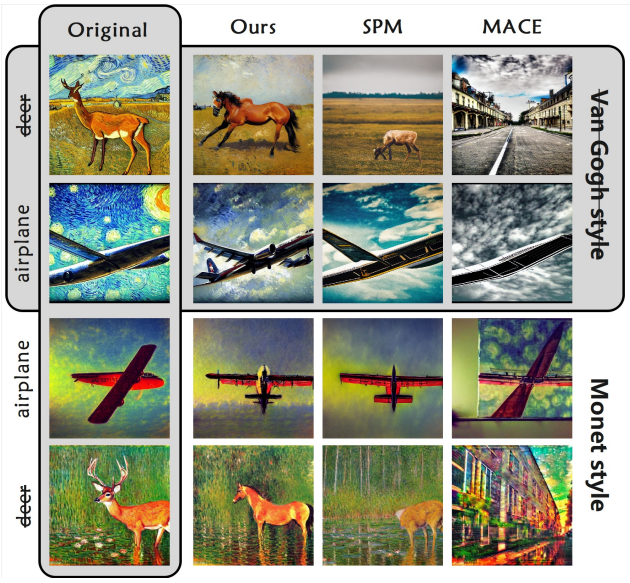

Figure 5. Samples of 'A photo of a {object} in {artist} style' after erasing *deer* and *Van Gogh*. MACE effectively removes *Van Gogh style* and *deer*, but severely damages *Monet style* and *airplane*. SPM successfully removes *Van Gogh style* but fails to eliminate *deer*. In contrast, our MapRoute successfully removes both *Van Gogh style* and *deer* while preserving others.

qualitative results for single-style erasure. It can be observed that while MACE maintains relatively stable erasure performance, it fails to adequately preserve other artistic styles. Both FMN and UCE do not completely erase the target artistic styles and partially impair the generation quality of other styles. SPM achieves a certain level of erasure effectiveness but still exhibits issues of concept erosion. In contrast, MapRoute achieves state-of-the-art performance across all metrics, demonstrating its superior erasure capability.

4.4. Mixed Concepts Erasure

To evaluate the effectiveness of the proposed method in mixed concept erasure tasks, this study constructs prompts by combining ten object categories from the CIFAR-10 dataset with three artistic styles—Van Gogh, Canaletto, and Monet—to systematically examine performance. For assessing the erasure and preservation capabilities of object concepts, we employ ACC_e and ACC_s same as section 4.1. As for artistic styles, the erasure and preservation effects are measured $CLIP_e$ and $CLIP_s$ same as section 4.3, respectively.

As shown in Figure 6, we compare the proposed method with MACE and SPM. The experimental results indicate that MACE achieves certain success in object concept erasure, as reflected by its relatively low ACC_e and high ACC_s, suggesting effective erasure of the target concepts. In terms

Table 3. **Quantitative results on artistic style erasure.** We use two CLIP scores to measure the model's ability to erase target concepts and preserve unrelated concepts.

Method	Brent Heighton			Brett Weston			Brett Whiteley		
	$\text{CLIP}_e \downarrow$	$\text{CLIP}_s \uparrow$	$H_a \uparrow$	$\text{CLIP}_e \downarrow$	$\text{CLIP}_s \uparrow$	$H_a \uparrow$	$\text{CLIP}_e \downarrow$	$\text{CLIP}_s \uparrow$	$H_a \uparrow$
FMN	22.68	24.22	1.54	24.08	24.97	0.89	21.40	24.50	3.10
UCE	22.07	25.51	3.44	22.33	23.64	1.29	21.22	25.18	3.96
SPM	20.39	24.99	4.60	19.40	25.03	5.63	19.76	24.99	5.23
MACE	15.95	22.77	6.82	**15.84**	22.75	6.91	19.08	22.80	3.72
Ours	**13.88**	26.09	**12.21**	17.78	26.09	**8.31**	**18.90**	26.17	**7.27**
SD v1.4	28.46	**26.18**	-2.28	25.47	**26.18**	0.71	25.18	**26.18**	1.00

Table 4. **Ablation study on crucial components.**

Configuration	Object Erasure			Artistic style Concepts	
	$\text{ACC}_e \downarrow$	$\text{ACC}_s \uparrow$	$\text{ACC}_g \downarrow$	$\text{CLIP}_e \downarrow$	$\text{CLIP}_s \uparrow$
Full Model	0.5	99.53	0.92	16.85	26.13
w/o L_{keep1}	5.44	19.49	5.87	13.21	11.43
w/o L_{keep2}	–	–	–	18.5	26.02
Two Map Layer	7.54	16.33	5.88	12.21	12.53
Four Map Layer	0.43	99.46	0.87	16.91	26.02

of artistic style processing, although MACE attains a good erasure effect (indicated by a high CLIP_e score), its CLIP_s value is significantly low, implying that the model struggles to generate images in other styles after removing the target style. This further corroborates its adapting to mixed concept scenarios.However, SPM does not demonstrate advantages across all evaluation metrics.

In contrast, MapRoute demonstrates superior performance in both concept erasure and preservation. In object concept erasure tasks, the ACC_e value remains consistently at 0.00 indicating complete elimination of the target concepts, while the ACC_s value stays at a high level confirming effective protection of non-target concepts. In artistic style processing, MapRoute achieves desirable results in both CLIP_e and CLIP_s metrics, demonstrating its ability to eliminate specified artistic styles while satisfactorily maintaining the model's capacity to generate images in other styles.

4.5. Ablation Study

To investigate the impact of key components in the experiment, we conducted ablation studies on critical structures of the model. Table 4 presents the results of object concepts erasure and artistic style concepts erasure after ablating specific components of the loss function.

The results indicate that removing L_{keep1} leads to a significant drop in ACC_s values in the entity concept erasure task, suggesting that the model suffers from catastrophic forgetting and fails to generate images effectively. When L_{keep2} is ablated, CLIP_e increase to varying degrees, indicating a decline in the model's ability to erase target concepts. A two-layer mapper fails to learn the necessary identity mapping, leading to unsuccessful image generation. While a four-layer mapper shows no significant improvement, it incurs a 49.97% parameter increase, which is not justified. Consequently, the three-layer configuration is adopted as it strikes an optimal balance between efficacy and efficiency. The results of additional ablation experiments are provided in the Appendix.

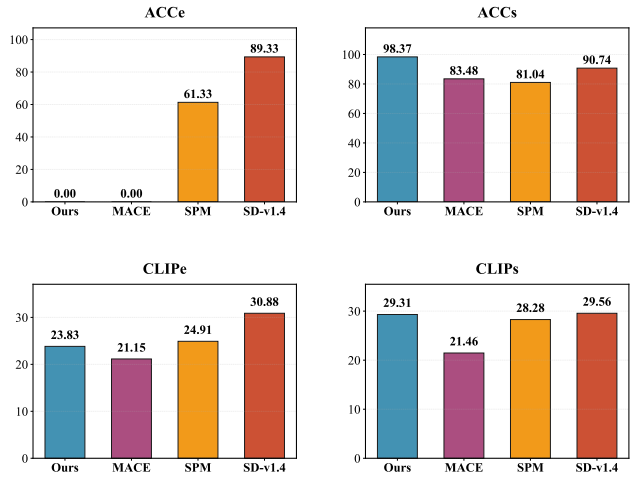

Figure 6. **Quantitative results on Mixed Concepts Erasure:** We construct prompts using the template "A photo of {object} in {artist} style" by combining the ten object categories from the CIFAR-10 dataset with three artistic styles: Van Gogh, Canaletto, and Monet. Images are generated under the conditions of erasing the concepts "dog", "truck", "deer", and "Van Gogh", respectively. Each prompt generate 50 images. The average values of ACCe, ACCs, CLIPe, and CLIPs are computed to evaluate the model's effectiveness in eliminating mixed concepts while preserving unrelated ones.

5. Conclusion

In this paper, we propose **MapRoute**, a lightweight, high-precision, and scalable framework for concept erasure in text-to-image diffusion models. MapRoute achieves precise intervention on target concepts by introducing a set of linear mapping modules immediately after the frozen pre-trained text encoder, without modifying any parameters of the diffusion backbone. During inference, the system dynamically activates the top-k Mappers most semantically relevant to the input prompt and applies their transformations in a serial manner. This mechanism not only significantly enhances the flexibility and composability of concept erasure but also effectively mitigates common challenges in multi-adapter deployment. Extensive experiments demonstrate that MapRoute achieves state-of-the-art performance in thoroughness, semantic consistency, and model fidelity across diverse concept types. Compared to existing methods, it exhibits advantages in both effectiveness and preservation of non-target content.

Acknowledgement

This research is supported by National Natural Science Foundation of China No. 62371156 and Shenzhen Science and Technology Program No. JCYJ20240813105132043.

References

[1] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063, 2023. 2

[2] Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). *Advances in Neural Information Processing Systems*, 37:84298–84328, 2024. 4

[3] Tom B Brown, Benjamin Mann, Melanie Ryder, Jared Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. 6

[4] Anh Bui, Trang Vu, Long Vuong, Trung Le, Paul Montague, Tamas Abraham, Junae Kim, and Dinh Phung. Fantastic targets for concept erasure in diffusion models and where to find them. *arXiv preprint arXiv:2501.18950*, 2025. 3

[5] Huiwen Chang, Han Zhang, Jarred Barber, A. J. Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. In *ICML*, 2023. 1

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[7] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2426–2436, 2023. 2, 5

[8] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120, 2024. 3, 5

[9] Feng Han, Kai Chen, Chao Gong, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Dumo: Dual encoder modulation network for precise concept erasure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3320–3328, 2025. 2

[10] Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. In *NeurIPS*, page 2, 2023. 6

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2, 3

[12] Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. In *European Conference on Computer Vision*, pages 360–376. Springer, 2024. 3, 5

[13] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2435, 2022. 1

[14] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 4, 5

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

[16] Byung Hyun Lee, Sungjin Lim, and Se Young Chun. Localized concept erasure for text-to-image diffusion models using training-free gated low-rank adaptation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18596–18606, 2025. 5

[17] Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. Safegen: Mitigating sexually explicit content generation in text-to-image models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 4807–4821, 2024. 2

[18] Yufan Liu, Jinyang An, Wanqian Zhang, Ming Li, Dayan Wu, Jingzi Gu, Zheng Lin, and Weiping Wang. Realera: Semantic-level concept erasure via neighbor-concept mining. *arXiv preprint arXiv:2410.09140*, 2024. 3

[19] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *CVPR*, pages 6430–6440, 2024. 3, 5

[20] Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7559–7568, 2024. 2, 3, 4, 5

[21] Zheling Meng, Bo Peng, Xiaochuan Jin, Yueming Lyu, Wei Wang, Jing Dong, and Tieniu Tan. Concept corrector: Erase concepts on the fly for text-to-image diffusion models. *arXiv preprint arXiv:2502.16368*, 2025. 2

[22] Pamela Mishkin, Lama Ahmad, Miles Brundage, Gretchen Krueger, and Girish Sastry. Dall· e 2 preview-risks and limitations. *Noudettu*, 28(2022):3, 2022. 2

[23] Minh Pham, Kelly O Marshall, Chinmay Hegde, and Niv Cohen. Robust concept erasure using task vectors. *arXiv preprint arXiv:2404.03631*, 2024. 2

[24] Alec Radford, Pranav Narasimhan, Tim Salimans, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 5

[25] 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. 1

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

[27] Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. Linear adversarial concept erasure. In *International Conference on Machine Learning*, pages 18400–18421. PMLR, 2022. 2

[28] 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. 1, 5, 7

[29] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1

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

[31] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 4

[32] Reza Shirkavand, Peiran Yu, Shangqian Gao, Gowthami Somepalli, Tom Goldstein, and Heng Huang. Efficient fine-tuning and concept suppression for pruned diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18619–18629, 2025. 2

[33] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6048–6058, 2023. 2

[34] Kartik Thakral, Tamar Glaser, Tal Hassner, Mayank Vatsa, and Richa Singh. Fine-grained erasure in text-to-image diffusion-based foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9121–9130, 2025. 3

[35] Jiahang Tu, Qian Feng, Jiahua Dong, Hanbin Zhao, Chao Zhang, Nicu Sebe, and Hui Qian. Ce-sdwv: Effective and efficient concept erasure for text-to-image diffusion models via a semantic-driven word vocabulary. *arXiv preprint arXiv:2501.15562*, 2025. 3

[36] Wikidata. Wikidata query service (sparql). https://query.wikidata.org/, 2025. Accessed: 2025-10-13. 5

[37] Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8496–8504, 2025. 2

[38] Yiwei Xie, Ping Liu, and Zheng Zhang. Erasing concepts, steering generations: A comprehensive survey of concept suppression. *arXiv preprint arXiv:2505.19398*, 2025. 2

[39] Yuyang Xue, Edward Moroshko, Feng Chen, Steven McDonagh, and Sotirios A Tsaftaris. Crce: Coreference-retention concept erasure in text-to-image diffusion models. *CoRR*, 2025. 3

[40] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6757–6767, 2023. 2

[41] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1755–1764, 2024. 5

[42] Xin Zhao, Xiaojun Chen, Yuexin Xuan, Zhendong Zhao, Xiaojun Jia, Xinfeng Li, and Xiaofeng Wang. Buster: Implanting semantic backdoor into text encoder to mitigate nsfw content generation. *arXiv preprint arXiv:2412.07249*, 2024. 2

[43] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2

[44] Hongguang Zhu, Yunchao Wei, Mengyu Wang, Siyu Jiao, Yan Fang, Jiannan Huang, and Yao Zhao. Sage: Exploring the boundaries of unsafe concept domain with semantic-augment erasing. *arXiv preprint arXiv:2506.09363*, 2025. 3

[45] Jingyu Zhu, Ruiqi Zhang, Licong Lin, and Song Mei. Choose your anchor wisely: Effective unlearning diffusion models via concept reconditioning. *CoRR*, 2024. 3