Mass Concept Erasure in Diffusion Models with Concept Hierarchy

Jiahang Tu[1], Ye Li[1], Yiming Wu[2], Hanbin Zhao*[1, 3], Chao Zhang[1], Hui Qian[1]

[1]College of Computer Science and Technology, Zhejiang University
[2]School of Computing and Data Science, The University of Hong Kong
[3]Zhejiang Key Laboratory of Intelligent High Speed UAV Technology
{tujiahang, yeli1, zhaohanbin, zczju, qianhui}@zju.edu.cn, yimingwu0@gmail.com

Abstract

The success of diffusion models has raised concerns about the generation of unsafe or harmful content, prompting concept erasure approaches that fine-tune modules to suppress specific concepts while preserving general generative capabilities. However, as the number of erased concepts grows, these methods often become inefficient and ineffective, since each concept requires a separate set of fine-tuned parameters and may degrade the overall generation quality. In this work, we propose a supertype-subtype concept hierarchy that organizes erased concepts into a parent–child structure. Each erased concept is treated as a child node, and semantically related concepts (*e.g.*, macaw, and bald eagle) are grouped under a shared parent node, referred to as a supertype concept (*e.g.*, bird). Rather than erasing concepts individually, we introduce an effective and efficient group-wise suppression method, where semantically similar concepts are grouped and erased jointly by sharing a single set of learnable parameters. During the erasure phase, standard diffusion regularization is applied to preserve denoising process in unmasked regions. To mitigate the degradation of supertype generation caused by excessive erasure of semantically related subtypes, we propose a novel method called **Su**pertype-**P**reserving **Lo**w-**R**ank **A**daptation (SuPLoRA), which encodes the supertype concept information in the frozen down-projection matrix and updates only the up-projection matrix during erasure. Theoretical analysis demonstrates the effectiveness of SuPLoRA in mitigating generation performance degradation. We construct a more challenging benchmark that requires simultaneous erasure of concepts across diverse domains, including celebrities, objects, and pornographic content. Comprehensive experiments demonstrate that our method achieves a superior balance between effective multi-concept erasure and the preservation of desirable generative performance.

Code — https://github.com/TtuHamg/SuPLoRA

1 Introduction

Recent advances in diffusion models (Song, Meng, and Ermon 2020; Ho, Jain, and Abbeel 2020; Tu et al. 2025b) have greatly improved text-to-image (T2I) generation, enabling users to produce high-quality and realistic images from simple text prompts. Tools like Stable Diffusion (SD) (Rombach et al. 2022; Podell et al. 2023), MidJourney (Midjourney 2024), and Flux (Labs 2024) highlight this capability. However, these advances raise major ethics (Jiang et al. 2023; Wang et al. 2024b), privacy (Mirsky and Lee 2021), and safety concerns (Wang et al. 2025b), as models often learn undesirable concepts, such as copyrighted materials, offensive content, and sensitive personal information, from unfiltered datasets (Rombach et al. 2022). Even with the high-cost data cleaning, diffusion models can still learn and generate unsafe content on filtered datasets (Schramowski et al. 2023).

To tackle this problem, various concept erasure methods have been proposed to **suppress the generation of concepts to be erased** while **preserving the model's capacity to generate general ones**. Early studies (Gandikota et al. 2023; Schramowski et al. 2023; Heng and Soh 2023; Fan et al. 2023; Li et al. 2024b; Yoon et al. 2024; Tu et al. 2025a; Feng et al. 2025a; Li et al. 2025a; Feng et al. 2024, 2025b) tune specific diffusion modules to erase single concepts. Recently, growing efforts have focused on mass concept erasure (Kumari et al. 2023; Zhang et al. 2024a; Zhao et al. 2024; Lu et al. 2024; Gandikota et al. 2024; Lyu et al. 2025), often via adapter tuning (Houlsby et al. 2019) or low-rank adaptation (LoRA) (Hu et al. 2022). These methods typically inject learnable parameters per concept and fine-tune them to suppress their generation.

However, these concept-wise erasure methods face two primary issues. Firstly, since the number of fine-tuned parameter sets grows linearly with the number of concepts to be erased, concept-wise erasure methods become inefficient as the number of erased concepts increases. This significantly increases storage overhead, making these methods impractical for real-world applications that require the erasure of a large number of concepts (OpenAI 2023). Secondly, since erasing a concept requires the model to suppress its learned generation patterns (Schramowski et al. 2023; Gandikota et al. 2023; Ho and Salimans 2022; Mei et al. 2025), repeatedly applying this process across multiple concepts inevitably degrades the model's generative capacity on general concepts (Lu et al. 2024; Zhao et al. 2024). For instance, in the celebrity erasure task, celebrities have distinct appearances, but they all belong to the supertype (Dai et al. 2023; Wang et al. 2025a) "person". When the model is instructed to forget an increasing number of celebrity identi-

*Corresponding author.

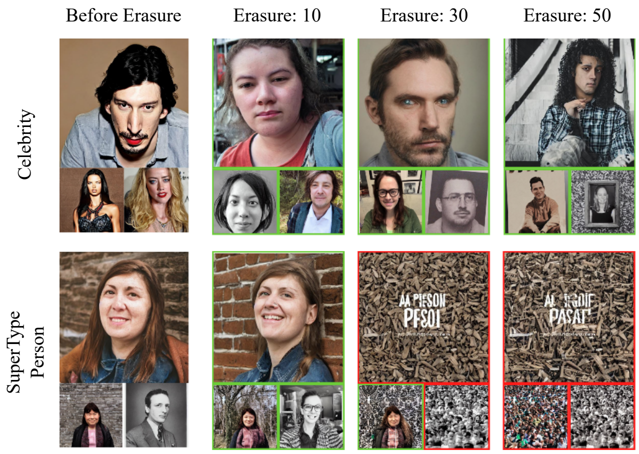

Figure 1: As more celebrities are erased, Stable Diffusion shows clear degradation in generating the supertype *person*. Results are shown for three representative celebrities ("*Adam Driver*", "*Adriana Lima*", "*Amber Heard*"). Green boxes in the first row indicate successful erasure. In the second row, green boxes indicate the generation capability of the supertype is preserved, while red boxes indicate failure.

ties, it will inevitably suppress visual features that are not only specific to those individuals but also essential for representing the supertype "person". As illustrated in Fig.1, increasing the number of erased celebrities results in noticeable degradation in the model's ability to generate the supertype "person".

To address these issues, we leverage the semantic relationship among concepts to be erased and construct a concept hierarchy (Dai et al. 2023; An et al. 2025) by organizing them into a parent–child structure. In this hierarchy, each erased concept is treated as a child node, and semantically related concepts, such as different species of birds (*e.g.*, jay, macaw, bald eagle), are grouped under a common parent node, which we define as a supertype concept (*e.g.*, bird). Instead of erasing concepts individually in previous methods, our supertype-subtype hierarchy enables group-wise erasure, where grouped child concepts are erased jointly in a more efficient manner. Specifically, we adopt MACE's (Lu et al. 2024) attention-based suppression to reduce the model focus on regions associated with the grouped concepts, and apply standard diffusion regularization to unmasked regions to preserve the model's denoising capability. To retain the generation capability for supertype concepts, we propose a **Su**pertype-**P**reserving **Lo**w-**R**ank **A**daptation (SuPLoRA), which designs the bases for representing the supertype concept subspace in the embedding space, initializes the down-projection matrix in SuPLoRA using orthogonal bases, and only trains the up-projection matrix. We further theoretically analyze the relationship between the update of down-projection matrix and the parameter of up-projection matrix in erasure setting, and prove that SuPLoRA mitigates performance degradation of supertype concepts. To evaluate the scalability and robustness of our method, we construct a more challenging benchmark than previous studies, which typically focus on erasing concepts from a single category. Our benchmark involves simultaneous erasure of concepts across multiple domains, including celebrities, objects, and pornographic content. Experimental results show that our method achieves a more favorable balance between the erasure of undesired concepts and the preservation of generative quality.

Our contributions are summarized as follows:

- We propose a concept hierarchy that organizes erased concepts into parent–child relationships. Based on this structure, we introduce a group-wise erasure strategy that jointly erases semantically related concepts under a shared supertype concept, improving performance efficiency over traditional concept-wise methods.
- We design SuPLoRA, a **Su**pertype-**P**reserving **Lo**w-**R**ank **A**daptation, to tackle the generation degradation of supertype concepts. Theoretical analysis ensures the effectiveness of SuPLoRA in mitigating generation performance degradation.
- We construct a more challenging benchmark spanning diverse concept, and demonstrate that our method achieves a superior trade-off between mass concept erasure and the preservation of generative quality.

2 Related Work

2.1 Inference-Time Intervention

Inference-time intervention methods aim to block undesired content by modifying the sampling process without changing model parameters. A common strategy (Schramowski et al. 2023; AUTOMATIC1111 2024) is to adjust classifier-free guidance (Ho and Salimans 2022; Wang et al. 2024a), as in Safe Latent Diffusion (Schramowski et al. 2023), which steers latent representations away from erased concepts. Another line of research (Tu et al. 2025a; Yoon et al. 2024; Li et al. 2024a; Wang et al. 2024c) focuses on manipulating the text embeddings used for conditioning. CE-SDWV (Tu et al. 2025a) constructs a semantic space representing the erased concepts and dynamically suppresses the corresponding semantic information that is hidden in the text embeddings. SAFREE (Yoon et al. 2024) applies subspace projection and adaptive re-attention to eliminate unsafe semantic directions in CLIP embeddings. While such methods avoid model fine-tuning and offer efficient erasure, they are fragile in practice, as the intervention can be bypassed simply by disabling the module during inference (Rando et al. 2022).

2.2 Tuning-based Erasure

To support safer model release, existing studies have explored fine-tuning methods to erase targeted concepts from pre-trained models. These approaches can be broadly categorized by the number of concepts they handle. Several works focus on single concept erasure (Gandikota et al. 2023; Heng and Soh 2023; Zhang et al. 2024a; Fan et al. 2023; Li et al. 2024b; Gao et al. 2025; Li et al. 2025b), aiming to remove one concept at a time. ESD (Gandikota et al. 2023) and AC (Kumari et al. 2023) align erased concepts with supertypes (*e.g.*, "grumpy cat" $\rightarrow$ "cat"), by fine-tuning cross-attention layers. SalUN (Fan et al. 2023) adopts a saliency-driven strategy that selectively updates parameters most associated with the target concept, minimizing side effects on general generation. In contrast, mass concept erasure (Zhao et al. 2024; Lu et al. 2024; Gandikota et al. 2024; Lyu et al. 2025; Kumari et al. 2023; Huang et al. 2023; Li et al. 2025a) targets multiple concepts simultaneously. For example, ConceptPrune (Chavhan, Li, and Hospedales 2024) prunes a union of "expert neurons" responsive to each concept, achieving erasure of ten object classes. MACE (Lu et al. 2024) leverages the LoRA fine-tuning technique, where each LoRA module is trained to erase a specific concept. However, most existing methods follow a concept-wise paradigm, in which each concept requires its own fine-tuned module. As the number of concepts increases, the total number of trainable parameters grows linearly, resulting in substantial storage overhead. This severely limits the scalability of these methods in real-world scenarios, where a large number of concepts may need to be erased simultaneously.

2.3 Preservation on Unerased Concepts

While existing methods effectively remove specific concepts, preserving generation for unerased concepts, particularly in mass concept erasure scenarios, remains a critical and underexplored challenge. To mitigate this, Selective Amnesia (SA) (Heng and Soh 2023) introduces a regularization term inspired by lifelong learning principles (Wang et al. 2022), encouraging the model to retain knowledge of unerased content. UCE (Gandikota et al. 2024) extends the TIME framework (Orgad, Kawar, and Belinkov 2023) by applying an auxiliary loss on a set of predefined preserved concepts and deriving a closed-form solution that balances erasure and retention objectives. SPM (Lyu et al. 2025) proposes an anchoring loss to protect distant, unrelated concepts during sequential erasure. Despite these efforts, a key limitation persists: as more concepts are erased, the model's ability to generate their shared supertype concept often degrades. In this work, we explicitly identify and address this issue by proposing SuPLoRA, a principled approach for preserving generation of unerased concepts.

3 Methodology

The objective of mass concept erasure is to remove a set of various concepts, while preserving the generative model's capability to produce high-fidelity outputs for general concepts. Let $\mathcal{M}$ be a pre-trained generative model with zero-shot capabilities over a broad set of general concepts $\mathcal{C}^g$. Let $\mathcal{C}^t = \{c_1^t, c_2^t, \ldots, c_N^t\}$ denote the set of concepts targeted for erasure, where $C^t \cap C^g = \emptyset$. The goal is to suppress the model's ability to generate content corresponding to each $c_i^t \in \mathcal{C}$, without degrading its generation quality on the general concepts. As the number of erased concepts increases, special attention must be given to preserving the generation capabilities of supertype concepts $\mathcal{C}^p = \{c_1^p, c_2^p, \ldots, c_K^p\}$. The erasure algorithm $\mathscr{E}$ aims to produce a modified model $\mathcal{M}' = \mathscr{F}(\mathcal{M}, \mathcal{C}^t, \mathcal{C}^g, \mathcal{C}^p)$ that modifies the mapping behavior as follows:

$$\begin{cases} f_{\mathcal{M}'}(\mathcal{T}_{\mathcal{C}^t}) \not\longrightarrow \mathcal{I}_{\mathcal{C}^t}, \\ f_{\mathcal{M}'}(\mathcal{T}_{\mathcal{C}^g}) \longrightarrow \mathcal{I}_{\mathcal{C}^g}, f_{\mathcal{M}'}(\mathcal{T}_{\mathcal{C}^p}) \longrightarrow \mathcal{I}_{\mathcal{C}^p}. \end{cases} \tag{1}$$

Here, $f_{\mathcal{M}'}(\mathcal{T})$ denotes the mapping from textual descriptions of concepts to image outputs $\mathcal{I}$, and $\not\longrightarrow$ indicates that the original mapping no longer holds after erasure.

3.1 Concept Hierarchy Construction

Previous approaches to mass concept erasure typically allocate a separate set of fine-tuning parameters for each target concept, treating each concept independently. However, many erased concepts exhibit semantic similarity, and overlooking these relationships leads to redundant parameter usage and storage overhead. To address this limitation, we propose to construct a concept hierarchy by exploiting the inherent semantic structure among concepts.

In our hierarchy, each concept to be erased is represented as a child node, while semantically related concepts are grouped under a common parent node, which we refer to as a supertype concept. We define the relationship as:

$$\mathcal{G}_j = \{c_i^t \in \mathcal{C}^t \mid g(c_i^t) = c_j^p\}, \tag{2}$$

where $g : \mathcal{C}^g \rightarrow \mathcal{C}^p$ s a mapping function that assigns each erased concept c_i^t to its corresponding supertype c_j^p. For example, as shown in Fig.2, different celebrities such as "Aaron Paul" and "Doris Day" are different erased concepts, but they share a common supertype "person". In this work, we construct a two-level concept hierarchy based on the supertype-subtype relationship among concepts. This hierarchical structure allows us to erase semantically similar concepts jointly by learning a shared set of parameters associated with their supertype. To build this hierarchy, we leverage the advanced semantic understanding of large language models (OpenAI 2023) (LLMs). Specifically, based on the semantic similarity among different concepts, we cluster multiple erased concepts into groups such that concepts within each group share a high-level semantic abstraction. Each group is then associated with a single supertype concept, identified using an LLM. This concept hierarchy lays the foundation for the group-wise suppression method described in the next section. Details on how supertype concepts are derived using LLMs are provided in App. B.1, and experiments on constructing a more complex multi-level concept hierarchy are provided in App. C.4.

3.2 Group-wise Suppression

To erase multiple concepts efficiently, we propose a group-wise suppression strategy based on the attention-based suppression method introduced in MACE (Lu et al. 2024). Inspired by MACE, our method minimizes the model's attention to concept-relevant regions to suppress undesired concepts. Unlike prior methods that treat each concept independently, we perform suppression at the supertype level, where semantically similar concepts are grouped and erased jointly by sharing a single set of learnable parameters. This design is motivated by our assumption that related concepts (*e.g.*, jay, macaw, bald eagle) can be represented within a shared semantic abstraction (*e.g.*, bird), and thus can be erased together through a unified parameter set. This reduces the total number of parameter sets from the number of erased concepts to the number of supertype groups, significantly improving parameter efficiency.

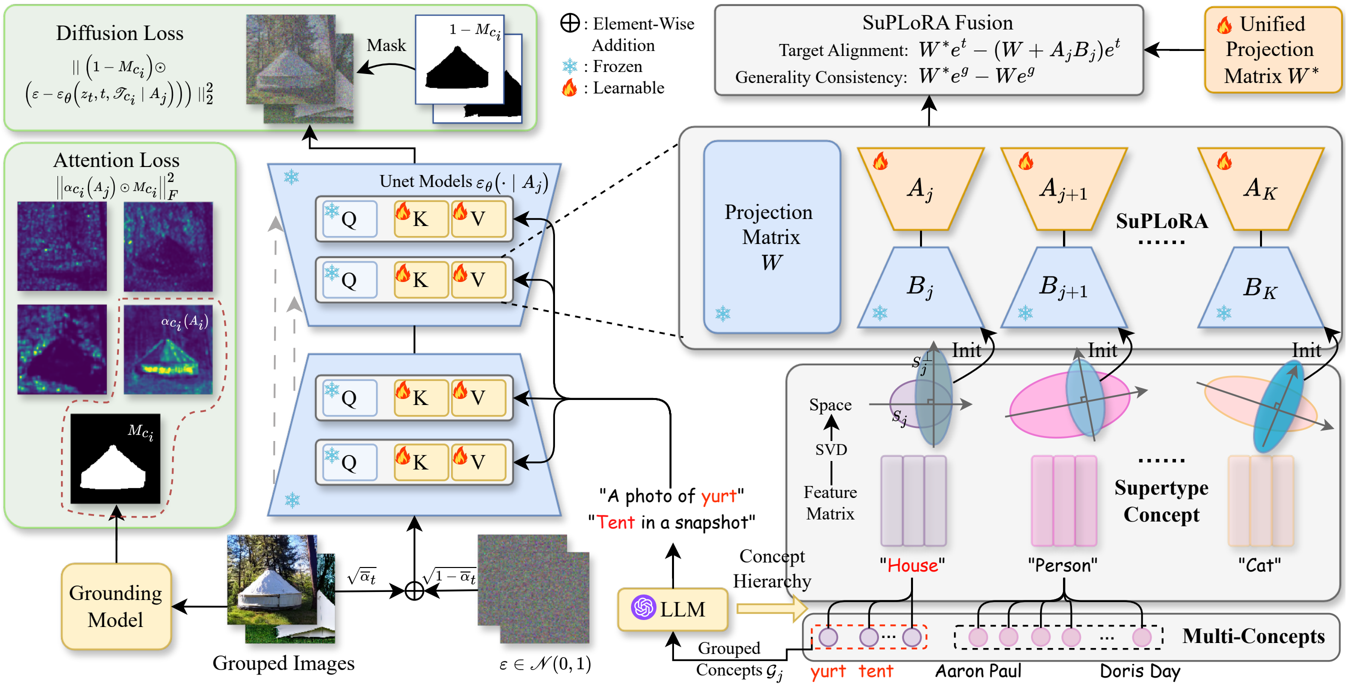

Figure 2: An overview of our mass concept erasure framework. We first construct a supertype–subtype concept hierarchy using LLMs (Sec.3.1). Leveraging this hierarchy, we introduce a group-wise suppression method that jointly erases grouped concepts via MACE's attention modulation, and apply diffusion regularization to preserve the model's denoising capability (Sec.3.2). To mitigate the degradation of supertype generation caused by excessive erasure of semantically related subtypes, we propose SuPLoRA, a supertype-preserving low-rank adaptation that freezes the down-projection matrix $\boldsymbol{B}_j$ (orthogonal to the supertype subspace) and updates only the up-projection matrix $\boldsymbol{A}_j$, as theoretically analyzed in Sec.3.3. Finally, we merge multiple SuPLoRA modules into a unified representation while preserving general generation ability (Sec. 3.3).

To achieve this, we utilize segmentation priors obtained from external grounding models (Liu et al. 2024) to identify concept-relevant areas $\boldsymbol{M}_{c_i}$ in the image. During training, the attention scores of concept-related tokens are selectively discouraged from attending to these areas, thereby weakening their spatial grounding. The erasure objective is defined as follows:

$$\mathcal{L}_{\text{attn}} = \mathbb{E}_{c_i \in \mathcal{G}_j, t, l} \left[\|\boldsymbol{\alpha}_{c_i}^{t,l}(\boldsymbol{A}_j) \odot \boldsymbol{M}_{c_i}\|_F^2 \right], \tag{3}$$

where $\boldsymbol{\alpha}_{c_i}^{t,l}(\boldsymbol{A}_j)$ denotes the attention map of concept token c_i at layer l and timestep t, whose values are influenced by the learnable parameters $\boldsymbol{A}_j$ injected into the attention projection layers.

While attention suppression effectively erase multiple concepts, it may also impair the model's ability to perform denoising, a core capability of diffusion-based generation. To address this, we apply standard diffusion training to the regions not associated with the erased concept:

$$\mathcal{L}_{\text{Diff}} = \mathbb{E}_{c_i \in \mathcal{G}_j, t, \boldsymbol{\epsilon}} \left[\|(1 - \boldsymbol{M}_{c_i}) \odot (\boldsymbol{\epsilon} - \epsilon_\theta(\boldsymbol{z_t}, t, \mathcal{T}_{c_i} | \boldsymbol{A}_j))\|_2^2 \right], \tag{4}$$

where $\epsilon_\theta(\cdot|\boldsymbol{A}_j)$ denotes the UNet models with learnable parameter $\boldsymbol{A}_j$. By fine-tuning the parameter $\boldsymbol{A}_j$, we aim to minimize the final loss:

$$\mathcal{L} = \mathcal{L}_{\text{attn}} + \lambda \mathcal{L}_{\text{Diff}}. \tag{5}$$

3.3 Supertype-Preserving Low-Rank Adaptation

Originally proposed as a parameter-efficient fine-tuning technique for large language models (OpenAI 2023) (LLMs), LoRA (Hu et al. 2022) has recently gained increasing attention in the vision community. It operates by injecting a pair of low-rank matrices, named as a down-projection matrix $\boldsymbol{B}$ and a up-projection matrix $\boldsymbol{A}$, into the weight layers of pre-trained models. This design enables the model to acquire task-specific knowledge while keeping the original weights largely unchanged. In group-wise erasure, the forward pass of a linear layer can be formulated as follows:

$$\boldsymbol{o}_j = \boldsymbol{W}\boldsymbol{h}_j + \boldsymbol{A}_j \boldsymbol{B}_j \boldsymbol{h}_j, \quad j \in [1, 2, \ldots, K], \tag{6}$$

where K denotes the number of supertype concept. Here, $\boldsymbol{h}_j$ and $\boldsymbol{o}_j$ represent the input and output of the linear layer, respectively. In the following paragraph, we analyze how to preserve the generative capability of a supertype concept during the erasure process.

We start by analyzing the relationship between fine-tuning SuPLoRA and directly tuning pre-trained weight $\boldsymbol{W}$. Suppose the j^{th} grouped concepts are targeted for erasure. Concept removal can be achieved either by directly modifying the pre-trained weights $\boldsymbol{W}$ or by learning the LoRA parameters $\boldsymbol{A}_j$ and $\boldsymbol{B}_j$.

If we opt to modify $\boldsymbol{W}$ directly, the gradient of the loss $\mathcal{L}$

with respect to $\boldsymbol{W}$ is computed via the chain rule:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{o}_j}\frac{\partial \boldsymbol{o}_j}{\partial \boldsymbol{W}} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{o}_j}\boldsymbol{h}_j^T. \tag{7}$$

With a learning rate α, the update to $\boldsymbol{W}$ is $\Delta \boldsymbol{W} = -\alpha \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}}$. Thus, the corresponding change in the erased matrix $\boldsymbol{W}' = \boldsymbol{W} + \boldsymbol{A}_j\boldsymbol{B}_j$ is:

$$\begin{aligned}\Delta_{\boldsymbol{W}}\boldsymbol{W}' =& [\boldsymbol{W} + \Delta\boldsymbol{W} + \boldsymbol{A}_j\boldsymbol{B}_j] - (\boldsymbol{W} + \boldsymbol{A}_j\boldsymbol{B}_j)\\ =& \Delta\boldsymbol{W} = -\alpha\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}} = -\alpha\frac{\partial \mathcal{L}}{\partial \boldsymbol{o}_j}\boldsymbol{h}_j^T.\end{aligned} \tag{8}$$

On the other hand, if we fine-tune $\boldsymbol{A}_j$ while keeping $\boldsymbol{W}$ and $\boldsymbol{B}_j$ fixed in Eq.6, its gradient is:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{A}_j} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{o}_j}\frac{\partial \boldsymbol{o}_j}{\partial \boldsymbol{A}_j} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{o}_j}\boldsymbol{h}_j^T\boldsymbol{B}_j^T. \tag{9}$$

Then, the update to $\boldsymbol{A}_j$ becomes $\Delta\boldsymbol{A}_j = -\alpha\frac{\partial \mathcal{L}}{\partial \boldsymbol{A}_j}$, and the change in the erased matrix $\boldsymbol{W}'$ is:

$$\begin{aligned}\Delta_{\boldsymbol{A}_j}\boldsymbol{W}' =& [\boldsymbol{W} + (\boldsymbol{A}_j + \Delta\boldsymbol{A}_j)\boldsymbol{B}_j] - (\boldsymbol{W} + \boldsymbol{A}_j\boldsymbol{B}_j)\\ =& \Delta\boldsymbol{A}_j\boldsymbol{B}_j = -\alpha\frac{\partial \mathcal{L}}{\partial \boldsymbol{A}_j}\boldsymbol{B}_j\\ =& \Delta_{\boldsymbol{W}}\boldsymbol{W}'\boldsymbol{B}_j^T\boldsymbol{B}_j. \quad \text{(based on Eq. 8)}\end{aligned} \tag{10}$$

In multi-domain concept erasure, Eq.10 reveals that ***fine-tuning $\boldsymbol{A}_j$ of the j^{th} grouped concepts is equivalent to modifying the pre-trained weight $\boldsymbol{W}$ within a subspace $\mathcal{S}_j^{\perp}$ defined by the projection matrix $\boldsymbol{B}_j^T\boldsymbol{B}_j$***. A similar view has been adopted in continual learning settings (Liang and Li 2024). Building upon the above formulation, ***if the subspace $\mathcal{S}_j^{\perp}$ defined by $\boldsymbol{B}_j$ is carefully designed to be orthogonal to the gradient subspace $\mathcal{S}_j$ associated with the learning of the supertype concept, then updates of $\boldsymbol{A}_j$ will lie in a direction orthogonal to that of the supertype concept gradients. Consequently, freezing $\boldsymbol{B}_j$ and fine-tuning $\boldsymbol{A}_j$ for concept erasure will not interfere with the model's generation ability of the supertype concept.*** This intuition aligns with findings in prior work on gradient orthogonality for task interference mitigation (Saha, Garg, and Roy 2021).

We then describe how to construct the subspace $\mathcal{S}_j^{\perp}$ and initialize $\boldsymbol{B}_j$. Prior studies (Liang and Li 2023; Saha, Garg, and Roy 2021) have demonstrated that the gradients of linear layers lie within the span of the input space. Leveraging this insight, SuPLoRA approximates the gradient subspace $\mathcal{S}_j$ for the j^{th} supertype concept using the input matrix $\boldsymbol{H}_{S_j}$ of pre-trained weight $\boldsymbol{W}$. When SuPLoRA is fine-tuned in the key and value projections of the cross-attention layer, $\boldsymbol{H}_{S_j}$ corresponds to the embeddings of the supertype concept descriptions. Typically, singular value decomposition (SVD) is applied to $\boldsymbol{H}_{S_j} = \boldsymbol{U}_j\boldsymbol{\Sigma}_j\boldsymbol{V}_j^T$, and the subspace $\mathcal{S}_j = \text{span}\{\boldsymbol{u}_{1,j}, \boldsymbol{u}_{2,j}, \ldots, \boldsymbol{u}_{r,j}\}$ is defined using the first r principle components. The orthogonal complement of this subspace, $\mathcal{S}_j^{\perp}$, can be computed via the null space of $\mathcal{S}_j$ or the projection operation in (Liang and Li 2024). Consequently, SuPLoRA sets $\boldsymbol{B}_i$ to the basis of $\mathcal{S}_j^{\perp}$ and fine-tunes only $\boldsymbol{A}_j$ to erase the j^{th} grouped concepts, while effectively preserving the generative capability of the supertype concept.

After obtaining K SuPLoRA modules from mass concept erasure, we adopt a knowledge distillation framework in (Lu et al. 2024) to obtain the final weight $\boldsymbol{W}^*$. The distillation objective is composed of two loss terms: 1) Target alignment loss aligns the output feature of $\boldsymbol{W}^*$ with those produced by individual SuPLoRA modules, ensuring the maintaining of erasure effects. 2) Generality consistency loss enforces feature-level consistency between the fused model and the base model when processing general concepts, thereby maintaining general generative capabilities.

$$\begin{aligned}\min_{\boldsymbol{W}^*}\mathbb{E}_{i,j}\|& \underbrace{\boldsymbol{W}^*\boldsymbol{e}_{j,i}^t - (\boldsymbol{W} + \boldsymbol{A}_j\boldsymbol{B}_j)\boldsymbol{e}_{j,i}^t}_{\text{target alignment}}\|_2^2\\ +\mathbb{E}_l& \underbrace{\|\boldsymbol{W}^*\boldsymbol{e}_l^g - \boldsymbol{W}\boldsymbol{e}_l^g\|_2^2}_{\text{generality consistency}}.\end{aligned} \tag{11}$$

Here, $\boldsymbol{A}_j\boldsymbol{B}_j$ is the j^{th} SuPLoRA module. $\boldsymbol{e}_{j,i}^t$, $\boldsymbol{e}_l^g$ are embeddings corresponding to erased concepts and general concepts, respectively.

4 Experiments

We conduct a challenging benchmark and comprehensively compare state-off-the-art baselines, including ESD-u (Gandikota et al. 2023), ESD-x (Gandikota et al. 2023), UCE (Gandikota et al. 2024), MACE (Lu et al. 2024), SPM (Lyu et al. 2025), CE-SDWV (Tu et al. 2025a), FMN (Zhang et al. 2024a), and SRS-ME (Zhao et al. 2024). Ablation studies are conducted to assess the contributions of key components in our approach. More experiments are provided in App. C.

4.1 Experimental Setup

Datasets. The erased concepts span three domains (Zhang et al. 2024b): objects (Fan et al. 2023), celebrities (Lu et al. 2024), and pornography (Schramowski et al. 2023). For the object domain, we select 30 target objects to be erased from ImageNet and retain 100 additional objects as the remaining concepts. In the celebrity domain, 30 target celebrities are chosen from the list provided by GIPHY Celebrity Detector (GCD) (Hasty et al. 2024), with 100 others preserved. For the pornography domain, we follow the definitions in (Lu et al. 2024) and select four target concepts. All concepts can be generated by SD v1.4 and classified using domain-specific classifiers. A complete list of the selected concepts is provided in App. A.

Implementation Details. We conduct experiments on SD v1.4 using the DDIM sampler (Song, Meng, and Ermon 2020) with 50 sampling steps. Following (Lu et al. 2024), erased concepts are augmented via GPT-4–generated descriptions (OpenAI 2023), and concept-relevant regions are localized using Grounded-SAM (Liu et al. 2024). Each SuPLoRA module is inserted into the key and value projections of the cross-attention layers and trained for 5 epochs with a learning rate of 0.0001. Since the input of the key or

Method	Is Mass	Erasure Effect			Domain-Specific		MS-COCO		Supertype	Efficiency	
		Cele Acc(↓)	Obj Acc(↓)	NN(↓)	Cele Acc(↑)	Obj Acc(↑)	FID(↓)	CLIP Score(↑)	CLIP Score (↑)	Storage/MB(↓)	Time/m(↓)
Methods that sacrifice generative performance for concept erasure											
ESD-x	✗	1.670%	15.40%	399	3.375%	52.50%	21.01	29.24	23.59	3379	2298
ESD-u	✗	0.000%	1.250%	59	0.500%	7.625%	34.59	25.21	22.05	3379	2166
FMN	✗	0.000%	0.000%	0	0.000%	0.000%	407.6	16.85	22.32	3379	24
CE-SDWV	✗	15.40%	16.25%	139	82.00%	61.25%	18.11	30.10	25.23	265	29
UCE	✓	9.870%	7.813%	163	73.62%	47.87%	18.51	29.80	24.81	3379	218
SRS-ME	✓	9.750%	9.000%	192	77.37%	52.12%	18.51	30.03	24.78	302	26
SPM	✓	10.00%	65.00%	639	78.50%	63.50%	21.15	30.59	26.00	218	20
MACE	✓	**6.250%**	9.167%	158	78.50%	50.63%	18.36	30.04	25.51	198	20
Ours	✓	7.500%	**4.167%**	**121**	**83.38%**	**65.00%**	**17.92**	**30.68**	**26.09**	**154**	**18**

Table 1: Assessment of Mass Concept Erasure. We evaluate both erasure effectiveness and the preservation of domain-specific, MS-COCO, and supertype concepts. NN denotes the explicit content detected by NudeNet on the I2P benchmark. Results with severely degraded generative quality are marked in red, while the best among acceptable methods are in **bold**. Our method achieves an efficient and effective balance between erasure and the preservation of desirable generation.

value projection matrix is from the text embedding, we employ the embeddings of supertype concept descriptions as $\boldsymbol{H}_{\mathcal{S}_j}$ and construct the subspace $\mathcal{S}_j$. The rank of SuPLoRA is set to 5, and the diffusion loss weight λ is 0.1. All baselines follow the settings in their original papers. Following (Lu et al. 2024; Lee et al. 2025), we use MS-COCO and unrelated concepts to construct the general concept set e^g. Additional implementation details are in App. B.

Evaluation Metric. The goal of mass concept erasure is to remove target concepts while maintaining general generative performance. We classify general concepts into three types: (1) domain-specific concepts retained within their domains; (2) supertype concepts as parent nodes in the hierarchy; and (3) MS-COCO concepts as general content unrelated to erased targets. To assess erasure and preservation, we use ViT-L/16 (88.06% top-1 accuracy) to classify images generated from erased and retained concepts. For the celebrity and pornography domains, we use the GCD classifier and NudeNet (Bedapudi 2019). We sample 10,000 MS-COCO prompts with minimal semantic overlap to generate images, then compute FID and CLIP Score for quality. Supertype preservation is evaluated via CLIP Scores of generated images. We also compare storage overhead and training time for efficiency.

4.2 Quantitative Analysis

In Tab. 1, we compare our method with both single-concept and mass-concept erasure approaches. Among single-concept methods, ESD-x, ESD-u, and FMN achieve low post-erasure classification accuracy for celebrities and objects, but suffer from severe degradation in generative performance (highlighted in red). For example, ESD-u yields a domain-specific celebrity accuracy of just 0.50% and a Supertype CLIP Score of 22.05, indicating that strong erasure comes at the cost of core generative capacity. CE-SDWV offers better generative preservation but weaker celebrity and obejct erasure. We next examine mass-concept erasure methods: UCE, SRS-ME, and MACE, which aim to balance erasure and generation. Among them, our method achieves the best trade-off, showing strong erasure effectiveness while preserving high domain-specific generation and competitive MS-COCO results. Thanks to the SuPLoRA design, our method also excels in supertype concept generation and reduces storage and training time compared to MACE.

4.3 Qualitative Comparison

From Fig. 3, our method and most baselines demonstrate effective qualitative erasure of target concepts. However, FMN achieves this at the cost of severely degrading generative ability, often producing noisy, unusable outputs. Single-concept methods (*e.g.*, ESD-x, ESD-u) also fail to consistently suppress targets like "beagle" or "Emma Roberts". Moreover, several methods struggle to preserve supertype concept generation, as shown in Fig. 1. Specifically, when erasing 30 concepts under the supertype "person" or "bird", models like SRS-ME, SPM and MACE lose the ability to generate coherent images. In contrast, our method not only removes target concepts effectively but also maintains high-quality generation for both domain-specific and supertype concepts. Additional results are provided in App. D.

4.4 Ablation Studies

Effect of SuPLoRA. Tab. 2 reports ablation results of SuPLoRA under different update configurations. None of the variants match the performance of our full method. Jointly updating both $\boldsymbol{A}_j$ and $\boldsymbol{B}_j$ leads to performance degradation, indicating that unconstrained updates can disrupt generative capacity. Freezing a randomly initialized $\boldsymbol{B}_j$ while updating only $\boldsymbol{A}_j$ performs better, likely due to reduced parameter flexibility mitigating such interference. SuPLoRA further improves results by explicitly constructing $\boldsymbol{B}_j$ to span the orthogonal complement of the supertype subspace, effectively preserving supertype generation. This design also enhances MS-COCO and domain-specific performance. All variants achieve strong erasure, with full results in App. C.

Effect of Key Components. In Tab. 3, we ablate the key components of our method. Removing the concept hierarchy and reverting to concept-wise erasure (*w/o* (1)) increases parameter cost from 7.11MB to 28.5MB, confirming its role in parameter efficiency. Excluding SuPLoRA (*w/o* (1)-(2)) minimally affects erasure but reduces domain-specific accuracy and supertype generation, underscoring its importance for generation preservation. Removing diffusion loss (*w/o*

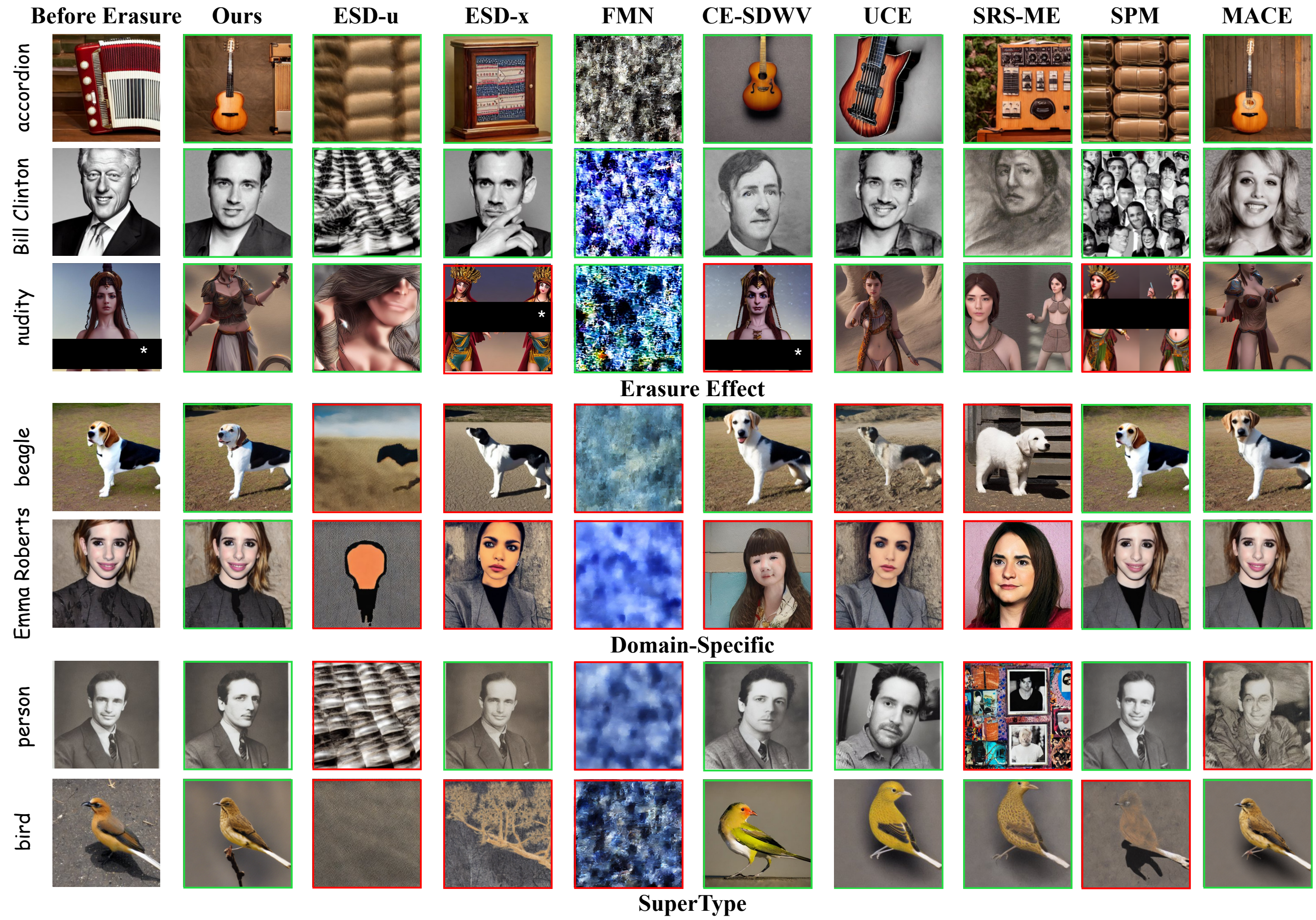

Figure 3: Qualitative comparison of mass concept erasure. Green boxes indicate successful removal of erased concepts and successful retention of preserved concepts. Red boxes highlight undesirable cases.

Configuration	Domain-Specific Cele/Obj Acc	MS-COCO FID	MS-COCO CLIP Score	Supertype CLIP Score
Default LoRA	79.12%/56.50%	18.18	30.18	25.19
Default LoRA, Freeze $\boldsymbol{B}_j$	81.12%/59.87%	18.13	30.65	26.08
SuPLoRA, Train $\boldsymbol{B}_j$	79.83%/57.01%	18.23	30.25	25.22
SuPLoRA	**83.38%/61.50%**	**17.94**	**30.66**	**26.21**

Table 2: Ablation study of comparing SuPLoRA and standard LoRA variants under varying configurations. "Freeze $\boldsymbol{B}_j$" denotes fine-tuning only $\boldsymbol{A}_j$ with a fixed $\boldsymbol{B}_j$, while "Train $\boldsymbol{B}_i$" updates both $\boldsymbol{A}_i$ and $\boldsymbol{B}_j$ jointly.

(1)-(3)) degrades all metrics except erasure, showing that denoising regularization is essential for output quality. Overall, our full method best balances effective erasure with the preservation of general and supertype-level generation.

5 Conclusion and Limitations

This paper tackles the challenge of erasing multiple concepts from diffusion models while preserving general concept generation. We build a semantic concept hierarchy and propose a group-wise suppression strategy that jointly erases related concepts under shared supertypes, improving parameter efficiency. We further introduce SuPLoRA, which freezes the down-projection and updates only the up-projection matrix, mitigating supertype degradation. However, our approach relies on shared supertype structures; when such overlap is limited, suppression may be less effective. Future work will explore adaptive, structure-independent erasure for better scalability.

Component	Erasure Acc	Domain Acc	MS-COCO FID/CLIP Score	Supertype CLIP Score	SuPLoRA Params
Ours	5.830%	**72.94%**	**17.94/30.66**	**26.21**	7.11MB
w/o (1)	6.040%	72.50%	18.01/30.63	25.98	28.5MB
w/o (1)-(2)	5.400%	67.81%	18.18/30.18	25.19	28.5MB
w/o (1)-(3)	**4.360%**	60.50%	18.97/29.39	24.99	28.5MB

Table 3: Ablation study on key components of our method. (1) denotes the concept hierarchy and group-wise erasure. (2) denotes the SuPLoRA. (3) denotes the diffusion loss during erasure process.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grant 62402430, 62206248 ,62476238, 62202431 and Zhejiang Provincial Natural Science Foundation of China under Grant LQN25F020008.

References

An, R.; Zeng, K.; Lu, M.; Yang, S.; Zhang, R.; Ji, H.; Zhang, Q.; Luo, Y.; Liang, H.; and Zhang, W. 2025. Concept-as-Tree: Synthetic Data is All You Need for VLM Personalization. *arXiv preprint arXiv:2503.12999*.

AUTOMATIC1111. 2024. Negative prompt. https://github.com/AUTOMATIC1111/stable-diffusion-webui/wiki/Negative-prompt. Accessed: 2025-11-24.

Bedapudi, P. 2019. Nudenet: Neural nets for nudity classification, detection and selective censoring.

Chavhan, R.; Li, D.; and Hospedales, T. 2024. Concept-Prune: Concept editing in diffusion models via skilled neuron pruning. *arXiv preprint arXiv:2405.19237*.

Dai, H.; Zhang, L.; Zhao, L.; Wu, Z.; Liu, Z.; Liu, D.; Yu, X.; Lyu, Y.; Li, C.; Liu, N.; et al. 2023. Hierarchical semantic tree concept whitening for interpretable image classification. *arXiv preprint arXiv:2307.04343*.

Fan, C.; Liu, J.; Zhang, Y.; Wong, E.; Wei, D.; and Liu, S. 2023. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*.

Feng, Q.; Tu, J.; Kang, M.; Zhao, H.; Zhang, C.; and Qian, H. 2025a. FG-OrIU: Towards Better Forgetting via Feature-Gradient Orthogonality for Incremental Unlearning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1957–1967.

Feng, Q.; Zhao, H.; Zhang, C.; Dong, J.; Ding, H.; Jiang, Y.-G.; and Qian, H. 2025b. Pectp: Parameter-efficient cross-task prompts for incremental vision transformer. *IEEE Transactions on Circuits and Systems for Video Technology*.

Feng, Q.; Zhou, D.-W.; Zhao, H.; Zhang, C.; Dong, J.; Dai, D.; and Qian, H. 2024. Lw2g: Learning whether to grow for prompt-based continual learning. *arXiv preprint arXiv:2409.18860*.

Gandikota, R.; Materzynska, J.; Fiotto-Kaufman, J.; and Bau, D. 2023. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2426–2436.

Gandikota, R.; Orgad, H.; Belinkov, Y.; Materzyńska, J.; and Bau, D. 2024. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5111–5120.

Gao, D.; Lu, S.; Zhou, W.; Chu, J.; Zhang, J.; Jia, M.; Zhang, B.; Fan, Z.; and Zhang, W. 2025. Eraseanything: Enabling concept erasure in rectified flow transformers. In *Forty-second International Conference on Machine Learning*.

Hasty, N.; Kroosh, I.; Voitekh, D.; et al. 2024. GIPHY Celebrity Detector. https://github.com/Giphy/celeb-detection-oss. Accessed: 2025-11-24.

Heng, A.; and Soh, H. 2023. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36: 17170–17194.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*, 2790–2799. PMLR.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huang, C.-P.; Chang, K.-P.; Tsai, C.-T.; Lai, Y.-H.; Yang, F.-E.; and Wang, Y.-C. F. 2023. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. *arXiv preprint arXiv:2311.17717*.

Jiang, H. H.; Brown, L.; Cheng, J.; Khan, M.; Gupta, A.; Workman, D.; Hanna, A.; Flowers, J.; and Gebru, T. 2023. AI Art and its Impact on Artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 363–374.

Kumari, N.; Zhang, B.; Wang, S.-Y.; Shechtman, E.; Zhang, R.; and Zhu, J.-Y. 2023. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22691–22702.

Labs, B. F. 2024. FLUX. https://github.com/black-forest-labs/flux. Accessed: 2025-11-24.

Lee, B. H.; Lim, S.; Lee, S.; Kang, D. U.; and Chun, S. Y. 2025. Concept pinpoint eraser for text-to-image diffusion models via residual attention gate. *arXiv preprint arXiv:2506.22806*.

Li, G.; Xiao, Y.; Ji, J.; Deng, K.; Hui, B.; Guo, L.; and Ma, X. 2025a. Sculpting Memory: Multi-Concept Forgetting in Diffusion Models via Dynamic Mask and Concept-Aware Optimization. *arXiv preprint arXiv:2504.09039*.

Li, L.; Lu, S.; Ren, Y.; and Kong, A. W.-K. 2025b. Set you straight: Auto-steering denoising trajectories to sidestep unwanted concepts. *arXiv preprint arXiv:2504.12782*.

Li, S.; van de Weijer, J.; Hu, T.; Khan, F. S.; Hou, Q.; Wang, Y.; and Yang, J. 2024a. Get What You Want, Not What You Don't: Image Content Suppression for Text-to-Image Diffusion Models. *arXiv preprint arXiv:2402.05375*.

Li, X.; Yang, Y.; Deng, J.; Yan, C.; Chen, Y.; Ji, X.; and Xu, W. 2024b. 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*, 4807–4821.

Liang, Y.-S.; and Li, W.-J. 2023. Adaptive Plasticity Improvement for Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7816–7825.

Liang, Y.-S.; and Li, W.-J. 2024. Inflora: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23638–23647.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 38–55. Springer.

Lu, S.; Wang, Z.; Li, L.; Liu, Y.; and Kong, A. W.-K. 2024. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6430–6440.

Lyu, M.; Yang, Y.; Hong, H.; Chen, H.; Jin, X.; He, Y.; Xue, H.; Han, J.; and Ding, G. 2025. 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*, 7559–7568.

Mei, L.; Guan, M.; Zheng, Y.; and Zhang, D. 2025. CoSketcher: Collaborative and Iterative Sketch Generation with LLMs under Linguistic and Spatial Control. In Christie, M.; Han, P.-H.; Lin, S.-S.; Pietroni, N.; Schneider, T.; Tsai, H.-R.; Wang, Y.-S.; and Zhang, E., eds., *Pacific Graphics Conference Papers, Posters, and Demos*. The Eurographics Association. ISBN 978-3-03868-295-0.

Midjourney. 2024. Midjourney Report. https://www.midjourney.com. Accessed: 2025-11-24.

Mirsky, Y.; and Lee, W. 2021. The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)*, 54(1): 1–41.

OpenAI. 2023. Gpt-4 technical report.

Orgad, H.; Kawar, B.; and Belinkov, Y. 2023. Editing implicit assumptions in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7053–7061.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Rando, J.; Paleka, D.; Lindner, D.; Heim, L.; and Tramèr, F. 2022. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Saha, G.; Garg, I.; and Roy, K. 2021. Gradient Projection Memory for Continual Learning. In *International Conference on Learning Representations*.

Schramowski, P.; Brack, M.; Deiseroth, B.; and Kersting, K. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22522–22531.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Tu, J.; Feng, Q.; Chen, C.; Dong, J.; Zhao, H.; Zhang, C.; and Qian, H. 2025a. CE-SDWV: Effective and Efficient Concept Erasure for Text-to-Image Diffusion Models via a Semantic-Driven Word Vocabulary. *arXiv preprint arXiv:2501.15562*.

Tu, J.; Ji, W.; Zhao, H.; Zhang, C.; Zimmermann, R.; and Qian, H. 2025b. Driveditfit: Fine-tuning diffusion transformers for autonomous driving data generation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(3): 1–29.

Wang, F.; Yin, H.; Dong, Y.-J.; Zhu, H.; Zhao, H.; Qian, H.; Li, C.; et al. 2024a. Belm: Bidirectional explicit linear multi-step sampler for exact inversion in diffusion models. *Advances in Neural Information Processing Systems*, 37: 46118–46159.

Wang, F.; Yin, H.; Qian, L.; Li, Y.; Zhuang, S.; Zhu, H.; Zhang, Y.; Tang, Y.; Zhang, C.; Zhao, H.; et al. 2025a. Unleashing High-Quality Image Generation in Diffusion Sampling Using Second-Order Levenberg-Marquardt-Langevin. *arXiv preprint arXiv:2505.24222*.

Wang, F.; Yin, H.; Zhuang, S.; Zhu, H.; Li, Y.; Qian, L.; Zhang, C.; Zhao, H.; Qian, H.; and Li, C. 2025b. Efficiently Access Diffusion Fisher: Within the Outer Product Span Space. *arXiv preprint arXiv:2505.23264*.

Wang, F.; Zhu, H.; Zhang, C.; Zhao, H.; and Qian, H. 2024b. GAD-PVI: A General Accelerated Dynamic-Weight Particle-Based Variational Inference Framework. In *AAAI*.

Wang, Y.; Li, O.; Mu, T.; Hao, Y.; Liu, K.; Wang, X.; and He, X. 2024c. Precise, Fast, and Low-cost Concept Erasure in Value Space: Orthogonal Complement Matters. *arXiv preprint arXiv:2412.06143*.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 139–149.

Yoon, J.; Yu, S.; Patil, V.; Yao, H.; and Bansal, M. 2024. Safree: Training-free and adaptive guard for safe text-to-image and video generation. *arXiv preprint arXiv:2410.12761*.

Zhang, G.; Wang, K.; Xu, X.; Wang, Z.; and Shi, H. 2024a. 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*, 1755–1764.

Zhang, Y.; Fan, C.; Zhang, Y.; Yao, Y.; Jia, J.; Liu, J.; Zhang, G.; Liu, G.; Kompella, R. R.; Liu, X.; et al. 2024b. UnlearnCanvas: Stylized Image Dataset for Enhanced Machine Unlearning Evaluation in Diffusion Models. *arXiv preprint arXiv:2402.11846*.

Zhao, M.; Zhang, L.; Zheng, T.; Kong, Y.; and Yin, B. 2024. Separable multi-concept erasure from diffusion models. *arXiv preprint arXiv:2402.05947*.