Novelty Assessment Report

Paper: SPEED: Scalable, Precise, and Efficient Concept Erasure for Diffusion Models

PDF URL: https://openreview.net/pdf?id=aoEtzdRkGh

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

Erasing concepts from large-scale text-to-image (T2I) diffusion models has become increasingly crucial due to the growing concerns over copyright infringement, offensive content, and privacy violations. In scalable applications, fine-tuning-based methods are time-consuming to precisely erase multiple target concepts, while real-time editing-based methods often degrade the generation quality of non-target concepts due to conflicting optimization objectives. To address this dilemma, we introduce SPEED, an efficient concept erasure approach that directly edits model parameters. SPEED searches for a null space, a model editing space where parameter updates do not affect non-target concepts, to achieve scalable and precise erasure. To facilitate accurate null space optimization, we incorporate three complementary strategies: Influence-based Prior Filtering (IPF) to selectively retain the most affected non-target concepts, Directed Prior Augmentation (DPA) to enrich the filtered retain set with semantically consistent variations, and Invariant Equality Constraints (IEC) to preserve key invariants during the T2I generation process. Extensive evaluations across multiple concept erasure tasks demonstrate that SPEED consistently outperforms existing methods in non-target preservation while achieving efficient and high-fidelity concept erasure, successfully erasing 100 concepts within only 5 seconds.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **concept erasure in text-to-image diffusion models**

A total of **50 papers** were analyzed and organized into a taxonomy with **19 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Core Erasure Mechanisms and Optimization Strategies**
- **Robustness and Adversarial Resilience**
- **Semantic Precision and Concept Disentanglement**
- **Scalability and Efficiency Optimization**
- **Specialized Erasure Contexts and Applications**
- **Evaluation Frameworks and Benchmarking**
- **Related Concept Manipulation and Editing**

Complete Taxonomy Tree

- concept erasure in text-to-image diffusion models Survey Taxonomy
- Core Erasure Mechanisms and Optimization Strategies
  - Fine-Tuning and Weight Modification Methods
  - Gradient-Based Fine-Tuning with Guidance (4 papers)
    - [2] Erasing concepts from diffusion models (Rohit Gandikota, 2023) View paper
    - [7] Erasing concepts from text-to-image diffusion models with few-shot unlearning (Takagi Tomohiro, 2024) View paper
    - [39] Towards Safe Self-Distillation of Internet-Scale Text-to-Image Diffusion Models (Kim, 2023) View paper
    - [43] ACE: Attentional Concept Erasure in Diffusion Models (Carter, 2025) View paper
  - Null Space and Direct Parameter Editing ★ (3 papers)
    - [0] SPEED: Scalable, Precise, and Efficient Concept Erasure for Diffusion Models (Anon et al., 2026) View paper
    - [40] Localized Concept Erasure for Text-to-Image Diffusion Models Using Training-Free Gated Low-Rank Adaptation (Byung Hyun Lee, 2025) View paper
    - [44] Editing Massive Concepts in Text-to-Image Diffusion Models (Xiong Tian-wei, 2024) View paper
  - Lightweight Adapter and Modular Erasure (2 papers)
    - [9] ICE: Intercede Concept Erasure in Text-to-Image Diffusion Models (Yizhou Lin, 2025) View paper
    - [12] Receler: Reliable Concept Erasing of Text-to-Image Diffusion Models via Lightweight Erasers (Chiâ□□pin Huang, 2023) View paper
  - Training-Free and Inference-Time Intervention
  - Text Embedding and Prompt Manipulation (4 papers)
    - [6] Unified concept editing in diffusion models (Rohit Gandikota, 2024) View paper
    - [22] Fantastic Targets for Concept Erasure in Diffusion Models and Where To Find Them (Bui Anh, 2025) View paper
    - [34] Get what you want, not what you don't: Image content suppression for text-to-image diffusion models (Li, 2024) View paper
    - [49] Semantic Surgery: Zero-Shot Concept Erasure in Diffusion Models (Liu Chengyu, 2025) View paper
  - Output-Side and Latent Intervention (2 papers)
    - [1] Concept Corrector: Erase concepts on the fly for text-to-image diffusion models (Peng Bo, 2025) View paper
    - [36] Mitigating inappropriate concepts in text-to-image generation with attention-guided Image editing. (Ji-Yeon Oh, 2025) View paper

  - Interpretability-Driven and Neuron-Level Erasure (2 papers)
  - [3] Sparse autoencoder as a zero-shot classifier for concept erasing in text-to-image diffusion models (Tian Zhi-hua, 2025) View paper
  - [21] A Single Neuron Works: Precise Concept Erasure in Text-to-Image Diffusion Models (He Qinqin, 2025) View paper
- Robustness and Adversarial Resilience
  - Adversarial Training for Robust Erasure (4 papers)
  - [17] Stereo: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models (Koushik Srivatsan, 2025) View paper
  - [18] Race: Robust adversarial concept erasure for secure text-to-image diffusion model (Chang-Hoon Kim, 2024) View paper
  - [20] Defensive Unlearning with Adversarial Training for Robust Concept Erasure in Diffusion Models (Chen Xin, 2024) View paper
  - [47] Rethinking Robust Adversarial Concept Erasure in Diffusion Models (Yin Qing-hong, 2025) View paper
  - Task Vector and Subspace-Based Robustness (2 papers)
  - [35] Robust concept erasure using task vectors (Pham, 2024) View paper
  - [37] SuMa: A Subspace Mapping Approach for Robust and Effective Concept Erasure in Text-to-Image Diffusion Models (Nguyen Kien, 2025) View paper
- Semantic Precision and Concept Disentanglement
  - Concept Combination and Multi-Concept Erasure (2 papers)
  - [4] Erasing Concept Combination from Text-to-Image Diffusion Model. (H Nie, 2025) View paper
  - [42] MACE: Mass Concept Erasure in Diffusion Models (Shilin Lu, 2024) View paper
  - Coreference and Semantic Graph Modeling (2 papers)
  - [19] CRCE: Coreference-Retention Concept Erasure in Text-to-Image Diffusion Models (Xue Yu-yang, 2025) View paper
  - [25] GrOCE:Graph-Guided Online Concept Erasure for Text-to-Image Diffusion Models (Ning Han, 2025) View paper
  - Concept Residue and Implicit Trace Removal (2 papers)
  - [16] Corer: Concept Residue Erasing in Text-to-Image Diffusion Models (Yufan Liu, 2025) View paper
  - [23] TRCE: Towards Reliable Malicious Concept Erasure in Text-to-Image Diffusion Models (Chen Rui-dong, 2025) View paper
- Scalability and Efficiency Optimization
  - Mass Concept Erasure and Batch Processing (3 papers)
  - [11] Reliable and efficient concept erasure of text-to-image diffusion models (Chao Gong, 2024) View paper
  - [24] ETCE: Efficient Two-Stage Concept Erasure for Text-to-Image Diffusion Models (Qingshuo Hu, 2025) View paper
  - [27] Ce-sdwv: Effective and efficient concept erasure for text-to-image diffusion models via a semantic-driven word vocabulary (Feng Qian, 2025) View paper
  - Online and Incremental Concept Erasure (2 papers)
  - [48] Continuous Concepts Removal in Text-to-image Diffusion Models (Han, 2024) View paper
  - [50] Now You See It, Now You Don't - Instant Concept Erasure for Safe Text-to-Image and Video Generation (Shristi Das Biswas, 2025) View paper
- Specialized Erasure Contexts and Applications
  - NSFW and Harmful Content Removal (3 papers)
  - [8] Comprehensive Assessment and Analysis for NSFW Content Erasure in Text-to-Image Diffusion Models (Chen Die, 2025) View paper
  - [14] All but One: Surgical Concept Erasing with Model Preservation in Text-to-Image Diffusion Models (SeungHoo Hong, 2023) View paper
  - [32] Comprehensive Evaluation and Analysis for NSFW Concept Erasure in Text-to-Image Diffusion Models (Chen Die, 2025) View paper
  - Copyright and Artistic Style Erasure (2 papers)
  - [5] Ablating concepts in text-to-image diffusion models (Nupur Kumari, 2023) View paper
  - [38] ACE: Anti-Editing Concept Erasure in Text-to-Image Models (Zihao Wang, 2025) View paper
  - Video and Temporal Concept Erasure (1 papers)
  - [30] VideoEraser: Concept Erasure in Text-to-Video Diffusion Models (Naen Xu, 2025) View paper
  - Concept Hiding and Reversible Erasure (2 papers)
  - [31] Hiding and recovering knowledge in text-to-image diffusion models via learnable prompts (Bui Anh, 2024) View paper
- Evaluation Frameworks and Benchmarking (4 papers)
  - [13] Erasing with Precision: Evaluating Specific Concept Erasure from Text-to-Image Generative Models (Takagi Tomohiro, 2025) View paper
  - [15] A comprehensive survey on concept erasure in text-to-image diffusion models (Kim, 2025) View paper
  - [45] Circumventing Concept Erasure Methods For Text-to-Image Generative Models (Pham, 2023) View paper
  - [46] Six-CD: Benchmarking Concept Removals for Text-to-image Diffusion Models (Jie Ren, 2025) View paper
- Related Concept Manipulation and Editing (5 papers)
  - [10] Editing implicit assumptions in text-to-image diffusion models (Hadas Orgad, 2023) View paper
  - [26] Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models (Shweta Mahajan, 2024) View paper
  - [28] A Comprehensive Survey on Visual Concept Mining in Text-to-image Diffusion Models (Li, 2025) View paper
  - [29] Scaling Concept With Text-Guided Diffusion Models (Huang Chao, 2024) View paper
  - [33] One-step diffusion for real-world image super-resolution via degradation removal and text prompts (Yaohui Guo, 2025) View paper

Narrative

Core task: concept erasure in text-to-image diffusion models. The field has organized itself around several complementary dimensions. Core Erasure Mechanisms and Optimization Strategies encompass foundational techniques—ranging from fine-tuning and weight modification (e.g., Erasing Concepts[2], Ablating Concepts[5]) to null-space projections and direct parameter editing—that directly alter model weights or internal representations to suppress unwanted concepts. Robustness and Adversarial Resilience addresses the challenge of adversarial prompts and jailbreaking attempts, ensuring that erasure remains effective under attack (e.g., Defensive Unlearning[20], Rethinking Robust Erasure[47]). Semantic Precision and Concept Disentanglement focuses on surgically removing target concepts without collateral damage to related semantics, while Scalability and Efficiency Optimization tackles the computational cost of editing large models or handling many concepts simultaneously (e.g., Editing Massive Concepts[44]). Specialized Erasure Contexts and Applications explore domain-specific needs such as NSFW content filtering (NSFW Assessment[8]) or video generation (VideoEraser[30]), and Evaluation Frameworks and Benchmarking provide standardized metrics (Precision Erasure Evaluation[13]) to compare methods.

Finally, Related Concept Manipulation and Editing extends erasure ideas to broader editing tasks, including style transfer and attribute modification.

A particularly active line of work contrasts fine-grained parameter surgery—where methods like Ablating Concepts[5] and Unified Concept Editing[6] carefully identify and modify specific weight subspaces—with more holistic optimization strategies that retrain or distill models under new constraints. Trade-offs between erasure precision, computational overhead, and robustness to adversarial recovery remain central open questions. SPEED[0] sits within the fine-tuning and direct parameter editing cluster, emphasizing null-space projections to achieve efficient, targeted erasure. Compared to neighbors like Localized Gated LoRA[40], which uses modular low-rank adapters for localized control, or Editing Massive Concepts[44], which scales erasure to hundreds of concepts, SPEED[0] prioritizes mathematical rigor in isolating concept directions within weight space, aiming for minimal side effects while maintaining computational efficiency. This positioning reflects a broader tension in the field between surgical precision and the practical demands of large-scale, robust deployment.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Localized Concept Erasure for Text-to-Image Diffusion Models Using Training-Free Gated Low-Rank Adaptation

Authors: Byung Hyun Lee, Sungjin Lim, Se Young Chun | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Fine-tuning based concept erasing has demonstrated promising results in preventing generation of harmful contents from text-to-image diffusion models by removing target concepts while preserving remaining concepts. To maintain the generation capability of diffusion models after concept erasure, it is necessary to remove only the image region containing the target concept when it locally appears in an image, leaving other regions intact. However, prior arts often compromise fidelity of the other ...

Relationship Analysis

Both papers belong to the Null Space and Direct Parameter Editing category, focusing on methods that identify and edit specific parameter subspaces to remove concepts without affecting others. While SPEED searches for a null space through SVD-based optimization with prior knowledge refinement techniques (IPF, DPA, IEC) to achieve scalable multi-concept erasure, GLoCE takes a training-free approach using gated low-rank adaptation matrices to enable localized concept erasure within specific image regions. The key difference is that SPEED focuses on global concept removal across entire images with null-space constraints, whereas GLoCE addresses the more specific problem of removing concepts from localized regions while preserving other areas of the same image.

2. Editing Massive Concepts in Text-to-Image Diffusion Models

Authors: Xiong Tian-wei, Wu Yue, Tianwei Xiong, Xie, Enze, et al. (11 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Text-to-image diffusion models suffer from the risk of generating outdated, copyrighted, incorrect, and biased content. While previous methods have mitigated the issues on a small scale, it is essential to handle them simultaneously in larger-scale real-world scenarios. We propose a two-stage method, Editing Massive Concepts In Diffusion Models (EMCID). The first stage performs memory optimization for each individual concept with dual self-distillation from text alignment loss and diffusion nois...

Relationship Analysis

Both papers belong to the Null Space and Direct Parameter Editing category, focusing on methods that identify and edit specific parameter subspaces to remove concepts without affecting others. They overlap in their use of closed-form solutions and null-space constraints to achieve concept erasure while preserving non-target concepts in text-to-image diffusion models. However, SPEED focuses on scalable erasure of up to 100 concepts through null-space optimization with Prior Knowledge Refinement techniques (IPF, DPA, IEC), while EMCID (Editing Massive Concepts in Diffusion Models) employs a two-stage approach with dual self-distillation and multi-layer editing to handle up to 1,000 concepts, targeting broader concept editing tasks including updating, rectifying, and debiasing beyond just erasure.

Contributions Analysis

Overall novelty summary. The paper proposes SPEED, a null-space constrained parameter editing method for concept erasure in text-to-image diffusion models. It resides in the 'Null Space and Direct Parameter Editing' leaf, which contains only three papers total, including SPEED itself. This leaf sits within the broader 'Fine-Tuning and Weight Modification Methods' branch, distinguishing itself from gradient-based iterative fine-tuning and lightweight adapter approaches. The sparse population of this specific leaf suggests that direct null-space optimization for concept erasure represents a relatively focused research direction within the larger field of 50 surveyed papers.

The taxonomy reveals that SPEED's immediate neighbors explore related parameter surgery techniques: one sibling addresses unified concept editing across multiple dimensions, while another employs localized gated adapters. Adjacent leaves contain gradient-based fine-tuning methods (four papers using negative guidance or distillation) and lightweight modular erasure approaches (two papers with separate adapter modules). The broader parent branch encompasses all weight modification strategies, contrasting with the sibling 'Training-Free and Inference-Time Intervention' branch that operates without parameter updates. SPEED's null-space formulation positions it at the intersection of mathematical rigor and direct weight editing, diverging from iterative optimization or modular decomposition strategies.

Among 30 candidates examined, the core null-space erasure contribution shows substantial prior work overlap, with 6 of 10 examined papers providing potentially refutable evidence. The Prior Knowledge Refinement framework (IPF, DPA, IEC techniques) appears more novel, with 0 refutable candidates among 10 examined. The efficiency claim of 350× speedup faces moderate overlap, with 2 of 10 candidates offering comparable scalability results. These statistics reflect a limited semantic search scope rather than exhaustive coverage. The null-space concept itself has established precedents, while the specific refinement strategies and their integration appear less explored in the examined literature.

Based on the top-30 semantic matches and taxonomy structure, SPEED occupies a sparsely populated but conceptually well-defined niche. The null-space formulation builds on recognized parameter editing principles, yet the three-component refinement framework introduces technical specificity not clearly anticipated by examined prior work. The analysis captures immediate semantic neighbors but cannot assess broader field coverage beyond the 50-paper taxonomy or alternative search strategies that might reveal additional overlaps.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: SPEED: Null-space constrained concept erasure method

Description: The authors propose SPEED, a method that formulates concept erasure as a null-space constrained optimization problem. By projecting parameter updates onto the null space of non-target concepts, SPEED achieves zero preservation error, enabling scalable and precise concept erasure without affecting non-target concepts while maintaining efficiency.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Machine unlearning via null space calibration

URL: View paper

Prior Art Analysis

Null Space Calibration[55] demonstrates that null-space constrained optimization for parameter editing was already applied to machine unlearning before SPEED. The candidate paper explicitly formulates parameter updates as projections onto null spaces to preserve non-target data, using the same mathematical framework (SVD decomposition, projection matrices, null space constraints) that SPEED claims as novel for concept erasure. Both papers solve optimization problems by projecting parameter updates onto null spaces defined by non-target samples, achieving zero preservation error through identical mathematical principles.

Evidence

Evidence 1 - **Rationale**: Both papers use null space projection to ensure parameter updates do not affect non-target samples. The candidate's prior work demonstrates this core principle was already established for parameter editing tasks. - **Original**: we search for the null space of prior knowledge, a model editing space where parameter updates do not affect the feature representations of non-target concepts. by projecting the model parameter updates for concept erasure onto such null space, speed can minimize the preservation error to zero - **Candidate**: to avoid over-unlearning, unsc constrains the unlearning process within a null space tailored to the remaining samples. this ensures unlearning does not negatively impact the model's performance on the remaining samples.

2. CaseEdit: Enhancing Localized Commonsense Reasoning via Null-Space Constrained Knowledge Editing in Small Parameter Language Models

URL: View paper

Brief Assessment

CaseEdit[58] applies null-space constrained optimization to knowledge editing in language models, not to concept erasure in diffusion models. The technical domains and objectives differ fundamentally.

3. Ace: Concept editing in diffusion models without performance degradation

URL: View paper

Prior Art Analysis

Ace Without Degradation[54] demonstrates that null-space constrained optimization for concept editing in diffusion models was proposed prior to SPEED. The candidate paper explicitly describes projecting parameter perturbations onto the null space of normal text representations to preserve them while erasing unsafe concepts. Both papers formulate concept erasure as a null-space constrained optimization problem, derive closed-form solutions, and apply singular value decomposition (SVD) to construct null-space projection matrices. The candidate paper's methodology directly anticipates SPEED's core technical approach of using null-space constraints to achieve zero preservation error.

Evidence

Evidence 1 - **Rationale**: Both papers describe projecting parameter updates onto the null space to achieve zero preservation error, showing the candidate's prior work on this exact mechanism. - **Original**: By projecting the model parameter updates for concept erasure onto such null space, speed can minimize the preservation error to zero without compromising erasure efficacy, thereby enabling scalable and precise concept erasure without affecting non-target concepts. - **Candidate**: by projecting parameter perturbations onto the null space of representations within llms, the representations could be unaffected by the perturbations. inspired by this, we propose ace, a novel editing method that extends null-space projection to diffusion models, enabling precise erasure of unsafe ...

Evidence 2 - **Rationale**: Both papers use identical SVD-based methods to construct the null-space projection matrix, showing the candidate's prior implementation of this technical approach. - **Original**: To projectΔonto null space, we apply singular value decomposition (svd) onc 0c⊤ 0 ∈r d0xd0 1 and have u,λ,u ⊤ =svd c0c⊤ , whereu∈r d0xd0 contains the singular vectors of c0c⊤ 0 , andλis a diagonal matrix of its singular values. The singular vectors inuw.r.t. zero singular values form an orthonor... - **Candidate**: following the conventional null space projection process (wang et al., 2021), we first perform singular value decomposition (svd) on t0 to obtain the left singular vector matrix u. Next, we remove the eigenvectors in u corresponding to zero eigenvalues, yielding ˆu. The projection matrix p is then c...

4. Mitigating Negative Interference in Multilingual Sequential Knowledge Editing through Null-Space Constraints

URL: View paper

Brief Assessment

Multilingual Null-Space[59] applies null-space constraints to multilingual sequential knowledge editing in LLMs, not to concept erasure in diffusion models. The technical domains and objectives are fundamentally different.

5. Null it out: Guarding protected attributes by iterative nullspace projection

URL: View paper

Prior Art Analysis

Null It Out[53] demonstrates that null-space constrained optimization for removing information from representations was proposed prior to SPEED. The candidate paper presents iterative null-space projection (INLP) as a method for removing information from neural representations by projecting onto the null space of linear classifiers. This directly challenges SPEED's novelty claim of being the first to formulate concept erasure as a null-space constrained optimization problem, as Null It Out[53] already established this approach for guarding protected attributes through null-space projection.

Evidence

Evidence 1 - **Rationale**: Both papers propose null-space projection methods for removing information from representations. The candidate explicitly describes this as a 'novel method' in their abstract, establishing prior work on null-space constrained approaches before SPEED. - **Original**: we propose speed, a scalable, precise, and efficient concept erasure method with null-space constrained model editing, capable of erasing 100 concepts in 5 seconds. - **Candidate**: we present iterative null-space projection (inlp), a novel method for removing information from neural representations. our method is based on repeated training of linear classifiers that predict a certain property we aim to remove, followed by projection of the representations on their null-space.

Evidence 2 - **Rationale**: Both papers describe using null-space projection to ensure parameter updates do not affect non-target information. The candidate's description of constructing projection matrices to nullify classifier predictions directly parallels SPEED's approach of projecting updates onto null space to achieve zero preservation error. - **Original**: To address the above limitations, we proposescalable, precise, andefficient concepterasure fordiffusion models (speed) (see fig. 1), an editing-based method incorporating null-space constraints. specifically, we search for thenull space of prior knowledge, a model editing space where parameter updat... - **Candidate**: given a set of vectors xi ∈rd and a set of corresponding discrete1 protected attributes zi ∈z, we seek a linear guarding function gthat remove the linear dependence between zand x. we begin with a high-level description of our approach. let cbe a trained linear classifier, parameterized by a matrix w...

Evidence 3 - **Rationale**: While SPEED introduces specific refinement strategies, the fundamental null-space projection mechanism for preserving non-target information was already established in the candidate paper, which describes the geometric interpretation and mathematical foundation of null-space projection for information removal. - **Original**: To facilitate accurate null space optimization, we incorporate three complementary strategies: influence-based prior filtering (ipf) to selectively retain the most affected non-target concepts, directed prior augmentation (dpa) to enrich the filtered retain set with semantically consistent variation... - **Candidate**: nullspace projection the linear interaction between w and a new test point xhas a simple geometric interpretation: xis projected on the subspace spanned by w's rows, and is classified according to the dot product between x and w's rows, which is proportional to the components of xin the direction of ...

Evidence 4 - **Rationale**: The candidate paper explicitly describes the method of finding null-space projection matrices to remove information while preserving non-target properties, which is the core concept that SPEED claims as novel. The candidate's method of calculating nullspace and constructing projection matrices predates SPEED's approach. - **Original**: we introduce speed, an efficient concept erasure approach that directly edits model parameters. speed searches for a null space, a model editing space where parameter updates do not affect non-target concepts, to achieve scalable and precise erasure. - **Candidate**: this suggests a simple method for rendering z linearly guarded for a set of vectors x: training a linear classifier that is parameterized by w0 to predict zfrom x, calculating its nullspace, finding the orthogonal projection matrix pn(w0) onto the nullspace, and using it to remove from x those compone...

6. Alphaedit: Null-space constrained knowledge editing for language models

URL: View paper

Prior Art Analysis

Alphaedit[51] demonstrates that null-space constrained optimization for model parameter editing was previously applied to language models. The candidate paper explicitly formulates parameter updates as a null-space constrained optimization problem, projecting perturbations onto the null space of preserved knowledge to achieve zero preservation error. This directly refutes the novelty claim that SPEED is the first to propose null-space constrained optimization for concept erasure, as Alphaedit[51] already applied this technique to knowledge editing in LLMs before SPEED's application to diffusion models.

Evidence

Evidence 1 - **Rationale**: Both papers use identical mathematical principles: projecting parameter updates onto the null space ensures zero disruption to preserved/non-target knowledge, demonstrating that this approach was already established in Alphaedit[51]. - **Original**: we search for the null space of prior knowledge, a model editing space where parameter updates do not affect the feature representations of non-target concepts. by projecting the model parameter updates for concept erasure onto such null space, speed can minimize the preservation error to zero - **Candidate**: if the perturbation Δ is projected into the null space of k0 (i.e., Δ′k0 = 0, where Δ′ denotes the projected perturbation), adding it to the parameters w results in: (w + Δ′)k0 = wk0 = v0. this implies that the projected Δ will not disrupt the key-value associations of the preserved knowledge

Evidence 2 - **Rationale**: While SPEED introduces specific strategies for null space construction, Alphaedit[51] already established the core null-space projection methodology using SVD and covariance matrices, showing the fundamental technique existed prior to SPEED. - **Original**: to facilitate accurate null space optimization, we incorporate three complementary strategies: influence-based prior filtering (ipf) to selectively retain the most affected non-target concepts, directed prior augmentation (dpa) to enrich the filtered retain set with semantically consistent variation... - **Candidate**: we adopt the null space of the non-central covariance matrix k0kt 0 ∈ rd0xd0 as a substitute to reduce computational complexity, as d0 is typically much smaller than 100, 000. this matrix's null space is equal to that of k0. following the existing methods for conducting null space projection (wang e...

7. EvoEdit: Evolving Null-space Alignment for Robust and Efficient Knowledge Editing

URL: View paper

Prior Art Analysis

EvoEdit[57] demonstrates that null-space constrained optimization for model parameter editing was previously proposed and applied to knowledge editing in language models. The candidate paper explicitly formulates parameter updates using null-space projection to preserve knowledge while making targeted edits, predating the original paper's application to concept erasure in diffusion models. Both papers use singular value decomposition to construct null-space projectors and apply them to constrain parameter updates, with EvoEdit[57] providing theoretical guarantees for this approach in sequential editing scenarios.

Evidence

Evidence 1 - **Rationale**: Both papers formulate the core approach as projecting parameter updates onto the null space to preserve non-target knowledge. EvoEdit[57] provides theoretical guarantees for this approach, demonstrating prior establishment of null-space constrained optimization for model editing. - **Original**: to address the above limitations, we proposescalable, precise, andefficient concepterasure fordiffusion models (speed), an editing-based method incorporating null-space constraints. specifically, we search for thenull space of prior knowledge, a model editing space where parameter updates do not aff... - **Candidate**: departing from conventional locate-then-edit paradigms, evoedit dynamically projects each new edit into thenull spaceof both preserved and previously edited knowledge before parameter integration. this approach theoretically guarantees output invariance for preserved knowledge, even after thousands ...

8. Jailbreaking prompt attack: A controllable adversarial attack against diffusion models

URL: View paper

Brief Assessment

Jailbreaking Prompt Attack[52] focuses on adversarial attacks to generate NSFW content by manipulating text embeddings, not on concept erasure or null-space constrained optimization for model editing.

9. VideoEraser: Concept Erasure in Text-to-Video Diffusion Models

URL: View paper

Brief Assessment

VideoEraser[30] focuses on text-to-video diffusion models with a two-stage training-free approach (SPEA and ARNG), while SPEED addresses text-to-image models through null-space constrained parameter editing. The technical approaches and application domains differ fundamentally.

10. CURE: Concept Unlearning via Orthogonal Representation Editing in Diffusion Models

URL: View paper

Prior Art Analysis

CURE[56] demonstrates that similar null-space constrained optimization approaches for concept erasure existed prior to SPEED. Both papers formulate concept erasure as a closed-form optimization problem that projects parameter updates onto orthogonal subspaces to preserve non-target concepts while removing target concepts. CURE[56] uses 'orthogonal projection' and 'singular value decomposition' to identify discriminative subspaces for concept removal, achieving erasure 'in only 2 seconds' through closed-form operations. This directly parallels SPEED's null-space constrained optimization that projects updates onto the null space of non-target concepts to achieve 'zero preservation error' and erase '100 concepts within only 5 seconds'. The core mathematical framework of using orthogonal projections in weight space for training-free concept erasure is shared between both methods.

Evidence

Evidence 1 - **Rationale**: Both papers propose training-free methods that operate directly in weight space for concept erasure with emphasis on efficiency and specificity. - **Original**: we propose speed, a scalable, precise, and efficient concept erasure method with null-space constrained model editing, capable of erasing 100 concepts in 5 seconds. - **Candidate**: we introduce cure, a training-free concept unlearning framework that operates directly in the weight space of pre-trained diffusion models, enabling fast, interpretable, and highly specific suppression of undesired concepts.

Evidence 2 - **Rationale**: Both methods use orthogonal projection techniques to isolate target concepts while preserving non-target concepts. SPEED uses null-space projection while CURE uses orthogonal projection via SVD, but the fundamental approach of projecting onto orthogonal subspaces is identical. - **Original**: we search for the null space of prior knowledge, a model editing space where parameter updates do not affect the feature representations of non-target concepts. By projecting the model parameter updates for concept erasure onto such null space, speed can minimize the preservation error to zero witho... - **Candidate**: at the core of our method is the spectral eraser, a closed-form, orthogonal projection module that identifies discriminative subspaces using singular value decomposition over token embeddings associated with the concepts to forget and retain. intuitively, the spectral eraser identifies and isolates ...

Evidence 3 - **Rationale**: Both papers emphasize achieving efficient concept erasure (SPEED: 5 seconds for 100 concepts, CURE: 2 seconds) through closed-form operations while preserving non-target generation capabilities. - **Original**: our extensive experiments show that speed consistently outperforms existing methods in prior preservation across various erasure tasks with minimal computational costs. - **Candidate**: All the processes above are in closed-form, guaranteeing extremely efficient erasure in only 2 seconds. Benchmarking against prior approaches, cure achieves a more efficient and thorough removal for targeted artistic styles, objects, identities, or explicit content, with minor damage to original g...

Contribution 2: Prior Knowledge Refinement framework with three complementary techniques

Description: The authors develop a framework called Prior Knowledge Refinement consisting of three techniques: Influence-based Prior Filtering (IPF) to select highly affected non-target concepts, Directed Prior Augmentation (DPA) to expand the retain set with semantically consistent variations, and Invariant Equality Constraints (IEC) to preserve key invariants during generation. These techniques work together to construct an accurate null space for effective model editing.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Pmet: Precise model editing in a transformer

URL: View paper

Brief Assessment

PMET[66] focuses on optimizing transformer component hidden states for precise model editing in language models, not on retain set refinement for diffusion model concept erasure. The technical domains and objectives are fundamentally different.

2. Mitigating the Language Mismatch and Repetition Issues in LLM-based Machine Translation via Model Editing

URL: View paper

Brief Assessment

Translation Model Editing[70] focuses on mitigating language mismatch and repetition errors in LLM-based machine translation by refining located FFN components, not on concept erasure in diffusion models with null-space constraints and retain set refinement.

3. History Matters: Temporal Knowledge Editing in Large Language Model

URL: View paper

Brief Assessment

History Matters[74] addresses temporal knowledge editing in language models, focusing on preserving historical knowledge when updating facts. The original paper focuses on concept erasure in diffusion models using null-space constraints with retain set refinement techniques (IPF, DPA, IEC). These are fundamentally different domains and technical approaches.

4. Assessing and post-processing black box large language models for knowledge editing

URL: View paper

Brief Assessment

Black Box Editing[72] focuses on knowledge editing for black-box LLMs in web applications, not on diffusion model concept erasure or retain set refinement for null-space constrained editing.

5. Targeted Angular Reversal of Weights (TARS) for Knowledge Removal in Large Language Models

URL: View paper

Brief Assessment

TARS[73] focuses on knowledge removal from LLMs through weight reversal in feed-forward networks, not on retain set refinement for diffusion model editing. The candidate addresses a fundamentally different problem domain (LLM knowledge removal vs. diffusion model concept erasure).

6. Which Retain Set Matters for LLM Unlearning? A Case Study on Entity Unlearning

URL: View paper

Brief Assessment

Retain Set Matters[68] focuses on analyzing which subsets of the retain set matter most for LLM unlearning (specifically entity unlearning), introducing the concept of syntactically similar neighbor sets. The ORIGINAL paper proposes a framework for concept erasure in diffusion models with three specific techniques (IPF, DPA, IEC) for refining retain sets in null-space constrained optimization. These are fundamentally different domains (LLM unlearning vs. diffusion model concept erasure) with different technical approaches.

7. Model Unlearning via Sparse Autoencoder Subspace Guided Projections

URL: View paper

Brief Assessment

Sparse Autoencoder Projections[71] focuses on machine unlearning in LLMs using SAE-guided subspace projections, not on concept erasure in diffusion models. The technical approaches differ fundamentally: the candidate uses SAE features for subspace construction in LLMs, while the original develops IPF, DPA, and IEC for refining retain sets in text-to-image diffusion models.

8. CURE: Concept Unlearning via Orthogonal Representation Editing in Diffusion Models

URL: View paper

Brief Assessment

CURE[56] does not describe techniques analogous to IPF, DPA, or IEC. CURE focuses on spectral regularization through singular vector modulation, not on filtering influenced concepts, augmenting retain sets with directed noise, or enforcing invariant equality constraints.

9. Large language model unlearning via embedding-corrupted prompts

URL: View paper

Brief Assessment

Embedding-Corrupted Prompts[69] focuses on LLM unlearning via prompt corruption in embedding space, not on diffusion model concept erasure or retain set refinement for model editing.

10. Wise: Rethinking the knowledge memory for lifelong model editing of large language models

URL: View paper

Brief Assessment

WISE[67] focuses on lifelong model editing for LLMs using dual parametric memory (main and side memory) with knowledge sharding and merging. The candidate does not address diffusion model concept erasure or null space optimization for retain sets, which are the core focus of the original paper's Prior Knowledge Refinement framework.

Contribution 3: Efficient multi-concept erasure achieving 350× speedup

Description: The authors demonstrate that SPEED achieves substantial computational efficiency, erasing 100 concepts in 5 seconds with a 350× speedup over competitive methods. This efficiency is achieved through closed-form optimization while maintaining superior prior preservation and erasure efficacy across various concept erasure tasks.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Leace: Perfect linear concept erasure in closed form

URL: View paper

Brief Assessment

Leace[60] focuses on linear concept erasure in closed form for fairness and interpretability, not on multi-concept erasure in diffusion models. The candidate addresses different technical domains (linear classifiers vs. diffusion model parameters) and application contexts.

2. Erasing Concepts, Steering Generations: A Comprehensive Survey of Concept Suppression

URL: View paper

Brief Assessment

Concept Suppression Survey[63] is a survey paper that reviews and categorizes existing concept erasure methods rather than proposing a novel erasure technique. It does not claim to introduce a new method achieving 350× speedup, but rather systematically reviews methods like SPEED and others in the literature.

3. MACE: Mass Concept Erasure in Diffusion Models

URL: View paper

Prior Art Analysis

MACE[42] demonstrates that it can successfully scale concept erasure up to 100 concepts using closed-form optimization combined with LoRA finetuning. The paper explicitly states that MACE achieves mass concept erasure through 'closed-form cross-attention refinement along with lora finetuning' and successfully handles up to 100 concepts. This directly challenges the novelty claim of SPEED being the first to achieve efficient multi-concept erasure (100 concepts) through closed-form optimization, as MACE[42] already demonstrated similar capabilities using closed-form methods for large-scale concept erasure.

Evidence

Evidence 1 - **Rationale**: Both papers claim to erase 100 concepts. MACE[42] explicitly uses 'closed-form cross-attention refinement' for mass concept erasure, demonstrating that closed-form optimization for large-scale multi-concept erasure was already established prior to SPEED's submission. - **Original**: speed searches for a null space, a model editing space where parameter updates do not affect non-target concepts, to achieve scalable and precise erasure... speed consistently outperforms existing methods in non-target preservation while achieving efficient and high-fidelity concept erasure, success... - **Candidate**: we introduce mace, a finetuning framework for the task of mass concept erasure. this task aims to prevent models from generating images that embody unwanted concepts when prompted... mace differs by successfully scaling the erasure scope up to 100 concepts... This is achieved by leveraging closed-fo...

4. Realera: Semantic-level concept erasure via neighbor-concept mining

URL: View paper

Brief Assessment

Realera[65] focuses on semantic-level erasure through neighbor-concept mining to address 'concept residue' issues, not on computational efficiency or closed-form optimization speedup claims.

5. Eraseanything: Enabling concept erasure in rectified flow transformers

URL: View paper

Brief Assessment

EraseAnything[61] focuses on concept erasure in rectified flow transformers (Flux models) using bi-level optimization with LoRA-based tuning, while the original paper addresses diffusion models with closed-form null-space optimization. The architectural contexts and optimization approaches differ fundamentally.

6. A comprehensive survey on concept erasure in text-to-image diffusion models

URL: View paper

Brief Assessment

Concept Erasure Survey[15] is a comprehensive survey paper that reviews existing concept erasure methods but does not present a novel erasure technique. It categorizes and discusses methods like MACE and UCE that achieve efficient erasure, but does not claim to be the first to propose efficient multi-concept erasure with closed-form optimization.

7. Editable concept bottleneck models

URL: View paper

Brief Assessment

Editable Bottleneck Models[64] focuses on editing concept bottleneck models for interpretability, not on efficient multi-concept erasure in diffusion models with closed-form optimization as described in the original paper.

8. Unified concept editing in diffusion models

URL: View paper

Prior Art Analysis

Unified Concept Editing[6] demonstrates that efficient multi-concept erasure with closed-form optimization was already achieved prior to the original paper. The candidate paper shows that their method can erase up to 100 concepts using closed-form solutions without training, achieving erasure in under 1 minute. This directly refutes the novelty claim of achieving 350× speedup through closed-form optimization, as the candidate paper already established this capability with similar or better efficiency (less than 1 minute for 100 concepts versus 5 seconds claimed in the original). Both papers use closed-form solutions to edit cross-attention weights, and the candidate explicitly states their method is faster than training-based approaches.

Evidence

Evidence 1 - **Rationale**: Both papers achieve efficient multi-concept erasure through closed-form optimization. The candidate demonstrates erasure of multiple concepts in under 1 minute, which is comparable to the 5 seconds claimed in the original paper for 100 concepts, suggesting the speedup achievement was not novel. - **Original**: speed immediately erases 100 concepts within 5 seconds, achieving new state-of-the-art (sota) performance with a350xspeedup over competitive methods. - **Candidate**: as a closed-form edit, modifying attention weights given the new keys and values mappings takes less than 1 minute. that enables efficient simultaneous editing of multiple concepts.

Evidence 2 - **Rationale**: The candidate paper already established closed-form parameter editing for scalable multi-concept erasure, demonstrating that hundreds of concepts can be edited efficiently without training, which challenges the novelty of the original paper's efficiency claims. - **Original**: speed searches for a null space, a model editing space where parameter updates do not affect non-target concepts, to achieve scalable and precise erasure. - **Candidate**: our method, called unified concept editing (uce), offers a fast and practical way to control model behavior post-training, filling the gaps where data curation might fall short. uce is a closed-form parameterediting method that enables the application of hundreds of editorial modifications within a sin...

Evidence 3 - **Rationale**: The candidate paper demonstrates multi-concept erasure with preservation capabilities using closed-form solutions, establishing prior work in efficient multi-concept erasure before the original paper's claimed contribution. - **Original**: we evaluate speed on three representative concept erasure tasks,i.e., fewconcept, multi-concept, and implicit concept erasure, where it consistently exhibits superior prior preservation across all erasure tasks. overall, our contributions can be summarized as follows: • we propose speed, a scalable,... - **Candidate**: our method can successfully erase multiple concepts while preserving the model's knowledge. we use the text embeddings of the artist names as our concepts ci to erase and a set of artists to preserve cj. as shown in figure 3, we are able to consistently erase multiple artistic styles, while other me...

Evidence 4 - **Rationale**: The candidate paper provides a closed-form solution for efficient parameter editing that preserves non-target concepts, demonstrating that the mathematical framework for efficient multi-concept erasure through closed-form optimization was already established. - **Original**: to address this dilemma, we introduce speed, an efficient concept erasure approach that directly edits model parameters. speed searches for a null space, a model editing space where parameter updates do not affect non-target concepts, to achieve scalable and precise erasure. - **Candidate**: w = (m∑ i=0 v∗ ict i + λwold)(m∑ i=0 cict i + λi)-1 the first term in the inverse matrix, m∑ i=0 cict i , is the covariance of the concept text embeddings being edited. as discussed in the appendix, we interpret the second term, an identity matrix, as matching the covariance of the large encyclope...

Evidence 5 - **Rationale**: The candidate paper demonstrates erasure of up to 100 concepts with closed-form methods, showing that the capability to erase 100 concepts efficiently was already achieved, refuting the novelty of the 350× speedup claim as a first achievement. - **Original**: extensive evaluations across multiple concept erasure tasks demonstrate that speed consistently outperforms existing methods in non-target preservation while achieving efficient and high-fidelity concept erasure, successfully erasing 100 concepts within only 5 seconds. - **Candidate**: our method can erase up to 100 artists simultaneously before damaging image fidelity and clip scores. after 50 erasures, the model's output for a given prompt and seed begins to change , as indicated by the lpips score, but remains aligned overall as evidenced by the clip score.

9. Erasebench: Understanding the ripple effects of concept erasure techniques

URL: View paper

Brief Assessment

EraseBench[62] focuses on evaluating the side effects and quality degradation of concept erasure techniques on non-target concepts, not on computational efficiency or speedup claims. The candidate does not address closed-form optimization or computational performance metrics that would challenge the original paper's novelty claim about achieving 350× speedup.

10. ACE: Attentional Concept Erasure in Diffusion Models

URL: View paper

Brief Assessment

ACE Attentional[43] focuses on attention-based concept erasure with adversarial fine-tuning, not closed-form optimization for multi-concept erasure. The candidate does not demonstrate prior work on achieving 350× speedup through closed-form methods as claimed in the original paper.

Appendix: Text Similarity Detection

Textual similarity detection checked 31 papers and found 5 similarity segment(s) across 4 paper(s).

The following **4 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Ace: Concept editing in diffusion models without performance degradation

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

2. Alphaedit: Null-space constrained knowledge editing for language models

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

3. Mitigating Negative Interference in Multilingual Sequential Knowledge Editing through Null-Space Constraints

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

4. EvoEdit: Evolving Null-space Alignment for Robust and Efficient Knowledge Editing

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] SPEED: Scalable, Precise, and Efficient Concept Erasure for Diffusion Models View paper
- [1] Concept Corrector: Erase concepts on the fly for text-to-image diffusion models View paper
- [2] Erasing concepts from diffusion models View paper
- [3] Sparse autoencoder as a zero-shot classifier for concept erasing in text-to-image diffusion models View paper
- [4] Erasing Concept Combination from Text-to-Image Diffusion Model. View paper
- [5] Ablating concepts in text-to-image diffusion models View paper
- [6] Unified concept editing in diffusion models View paper
- [7] Erasing concepts from text-to-image diffusion models with few-shot unlearning View paper
- [8] Comprehensive Assessment and Analysis for NSFW Content Erasure in Text-to-Image Diffusion Models View paper
- [9] ICE: Intercede Concept Erasure in Text-to-Image Diffusion Models View paper
- [10] Editing implicit assumptions in text-to-image diffusion models View paper
- [11] Reliable and efficient concept erasure of text-to-image diffusion models View paper
- [12] Receler: Reliable Concept Erasing of Text-to-Image Diffusion Models via Lightweight Erasers View paper
- [13] Erasing with Precision: Evaluating Specific Concept Erasure from Text-to-Image Generative Models View paper
- [14] All but One: Surgical Concept Erasing with Model Preservation in Text-to-Image Diffusion Models View paper
- [15] A comprehensive survey on concept erasure in text-to-image diffusion models View paper
- [16] Corer: Concept Residue Erasing in Text-to-Image Diffusion Models View paper
- [17] Stereo: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models View paper
- [18] Race: Robust adversarial concept erasure for secure text-to-image diffusion model View paper
- [19] CRCE: Coreference-Retention Concept Erasure in Text-to-Image Diffusion Models View paper
- [20] Defensive Unlearning with Adversarial Training for Robust Concept Erasure in Diffusion Models View paper
- [21] A Single Neuron Works: Precise Concept Erasure in Text-to-Image Diffusion Models View paper
- [22] Fantastic Targets for Concept Erasure in Diffusion Models and Where To Find Them View paper
- [23] TRCE: Towards Reliable Malicious Concept Erasure in Text-to-Image Diffusion Models View paper
- [24] ETCE: Efficient Two-Stage Concept Erasure for Text-to-Image Diffusion Models View paper
- [25] GrOCE:Graph-Guided Online Concept Erasure for Text-to-Image Diffusion Models View paper
- [26] Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models View paper
- [27] Ce-sdwv: Effective and efficient concept erasure for text-to-image diffusion models via a semantic-driven word vocabulary View paper
- [28] A Comprehensive Survey on Visual Concept Mining in Text-to-image Diffusion Models View paper
- [29] Scaling Concept With Text-Guided Diffusion Models View paper
- [30] VideoEraser: Concept Erasure in Text-to-Video Diffusion Models View paper
- [31] Hiding and recovering knowledge in text-to-image diffusion models via learnable prompts View paper
- [32] Comprehensive Evaluation and Analysis for NSFW Concept Erasure in Text-to-Image Diffusion Models View paper
- [33] One-step diffusion for real-world image super-resolution via degradation removal and text prompts View paper
- [34] Get what you want, not what you don't: Image content suppression for text-to-image diffusion models View paper
- [35] Robust concept erasure using task vectors View paper
- [36] Mitigating inappropriate concepts in text-to-image generation with attention-guided Image editing. View paper
- [37] SuMa: A Subspace Mapping Approach for Robust and Effective Concept Erasure in Text-to-Image Diffusion Models View paper
- [38] ACE: Anti-Editing Concept Erasure in Text-to-Image Models View paper
- [39] Towards Safe Self-Distillation of Internet-Scale Text-to-Image Diffusion Models View paper
- [40] Localized Concept Erasure for Text-to-Image Diffusion Models Using Training-Free Gated Low-Rank Adaptation View paper
- [41] Removing Undesirable Concepts in Text-to-Image Generative Models with Learnable Prompts View paper
- [42] MACE: Mass Concept Erasure in Diffusion Models View paper
- [43] ACE: Attentional Concept Erasure in Diffusion Models View paper
- [44] Editing Massive Concepts in Text-to-Image Diffusion Models View paper
- [45] Circumventing Concept Erasure Methods For Text-to-Image Generative Models View paper
- [46] Six-CD: Benchmarking Concept Removals for Text-to-image Diffusion Models View paper
- [47] Rethinking Robust Adversarial Concept Erasure in Diffusion Models View paper
- [48] Continuous Concepts Removal in Text-to-image Diffusion Models View paper

- [49] Semantic Surgery: Zero-Shot Concept Erasure in Diffusion Models View paper
- [50] Now You See It, Now You Don't - Instant Concept Erasure for Safe Text-to-Image and Video Generation View paper
- [51] Alphaedit: Null-space constrained knowledge editing for language models View paper
- [52] Jailbreaking prompt attack: A controllable adversarial attack against diffusion models View paper
- [53] Null it out: Guarding protected attributes by iterative nullspace projection View paper
- [54] Ace: Concept editing in diffusion models without performance degradation View paper
- [55] Machine unlearning via null space calibration View paper
- [56] CURE: Concept Unlearning via Orthogonal Representation Editing in Diffusion Models View paper
- [57] EvoEdit: Evolving Null-space Alignment for Robust and Efficient Knowledge Editing View paper
- [58] CaseEdit: Enhancing Localized Commonsense Reasoning via Null-Space Constrained Knowledge Editing in Small Parameter Language Models View paper
- [59] Mitigating Negative Interference in Multilingual Sequential Knowledge Editing through Null-Space Constraints View paper
- [60] Leace: Perfect linear concept erasure in closed form View paper
- [61] Eraseanything: Enabling concept erasure in rectified flow transformers View paper
- [62] Erasebench: Understanding the ripple effects of concept erasure techniques View paper
- [63] Erasing Concepts, Steering Generations: A Comprehensive Survey of Concept Suppression View paper
- [64] Editable concept bottleneck models View paper
- [65] Realera: Semantic-level concept erasure via neighbor-concept mining View paper
- [66] Pmet: Precise model editing in a transformer View paper
- [67] Wise: Rethinking the knowledge memory for lifelong model editing of large language models View paper
- [68] Which Retain Set Matters for LLM Unlearning? A Case Study on Entity Unlearning View paper
- [69] Large language model unlearning via embedding-corrupted prompts View paper
- [70] Mitigating the Language Mismatch and Repetition Issues in LLM-based Machine Translation via Model Editing View paper
- [71] Model Unlearning via Sparse Autoencoder Subspace Guided Projections View paper
- [72] Assessing and post-processing black box large language models for knowledge editing View paper
- [73] Targeted Angular Reversal of Weights (TARS) for Knowledge Removal in Large Language Models View paper
- [74] History Matters: Temporal Knowledge Editing in Large Language Model View paper