

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

Yiwei Xie, Ping Liu[✉], Zheng Zhang[✉]

Abstract—Text-to-Image (T2I) models have demonstrated impressive capabilities in generating high-quality and diverse visual content from natural language prompts. However, uncontrolled reproduction of sensitive, copyrighted, or harmful imagery poses serious ethical, legal, and safety challenges. To address these concerns, the concept erasure paradigm has emerged as a promising direction, enabling the selective removal of specific semantic concepts from generative models while preserving their overall utility. This survey provides a comprehensive overview and in-depth synthesis of concept erasure techniques in T2I diffusion models. We systematically categorize existing approaches along three key dimensions: intervention level, which identifies specific model components targeted for concept removal; optimization structure, referring to the algorithmic strategies employed to achieve suppression; and semantic scope, concerning the complexity and nature of the concepts addressed. This multi-dimensional taxonomy enables clear, structured comparisons across diverse methodologies, highlighting fundamental trade-offs between erasure specificity, generalization, and computational complexity. We further discuss current evaluation benchmarks, standardized metrics, and practical datasets, emphasizing gaps that limit comprehensive assessment, particularly regarding robustness and practical effectiveness. Finally, we outline major challenges and promising future directions, including disentanglement of concept representations, adaptive and incremental erasure strategies, adversarial robustness, and new generative architectures. This survey aims to guide researchers toward safer, more ethically aligned generative models, providing foundational knowledge and actionable recommendations to advance responsible development in generative AI.

Index Terms—Text-to-Image Synthesis, Concept Suppression, Controllable Content Generation

I. INTRODUCTION

TEXT-TO-IMAGE (T2I) generation, driven by deep generative models such as GANs [1]–[3], VAEs [4]–[6], and diffusion models [7]–[9], has achieved remarkable success in synthesizing realistic and diverse images from natural language prompts [10]–[12]. These models are now widely deployed in creative industries, personalized content generation, and interactive media applications. Despite these advances, current T2I models still offer limited control over the inclusion or exclusion of sensitive semantic concepts, especially in open-ended or user-driven scenarios. As a result, they may inadvertently generate unsafe or inappropriate content, including

pornographic scenes, violent imagery, or biased representations [13]–[16]. This limitation presents a critical challenge for the safe and trustworthy deployment of generative systems in open-world settings.

Despite growing efforts in content alignment, the limited controllability of current T2I models still presents potential safety, legal, and ethical risks in real-world deployment. These systems may inadvertently generate content involving nudity, graphic violence, or copyrighted elements, even when such concepts are only implicitly embedded in user prompts. Moreover, they can amplify societal biases related to gender, race, or age, leading to skewed or discriminatory visual outputs. Such risks compromise not only fairness and user trust, but also raise serious concerns regarding legal compliance and platform accountability. Therefore, there is an urgent need for techniques that can selectively erase unsafe or undesirable concepts from generative models while maintaining their expressive capacity and overall utility.

Concept erasure methods aim to suppress the generation of specific semantic concepts, such as nudity, violence, or copyrighted elements, without retraining the model or degrading its general expressive capacity. These techniques intervene at different levels of the generation pipeline (e.g., text encoder, attention layers, diffusion trajectories) to decouple, override, or neutralize internal representations associated with the target concept. This allows the model to avoid reproducing sensitive content even when such concepts are explicitly invoked in the input prompt. Figure 1 illustrates examples of representative concepts that can be effectively erased through such interventions. These include a diverse range of categories, such as artistic styles, physical objects, celebrity identify, and sensitive content like nudity. By suppressing these concepts at various stages of the model’s representation, concept erasure allows for fine-grained control over the generated output, ensuring that sensitive or inappropriate content is not produced. Although often compared to machine unlearning [17]–[19], concept erasure serves a fundamentally different role: rather than retroactively removing training data influence, it enables fine-grained suppression of semantic concepts at inference time, offering a scalable and modular approach to model-level safety alignment. Machine unlearning addresses privacy or copyright concerns by removing the influence of specific training data [20], often requiring model re-optimization or parameter reinitialization to simulate “never learned” behavior. In contrast, concept erasure assumes that semantic concepts can be localized within distinct subspaces of the model’s

Yiwei Xie and Zheng Zhang are with Huazhong University of Science and Technology, China (email: {yiweixie, leaf}@hust.edu.cn)

Ping Liu is with University of Nevada, Reno, NV, USA (email: pino.pingliu@gmail.com)

[✉] denotes the co-corresponding author.

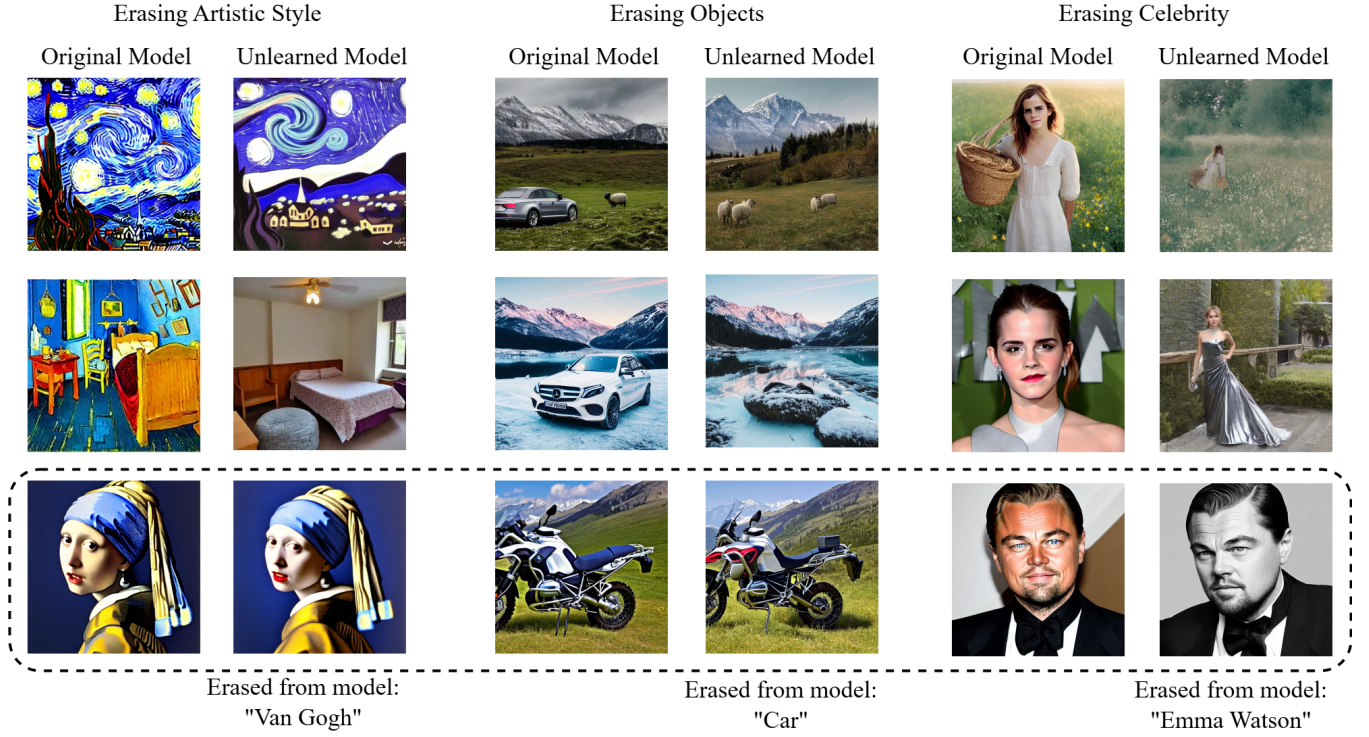


Fig. 1. Concept erasure using ESD. The top two rows show image generations from the original model (left) and the unlearned model (right) across three erasure scenarios: (1) artistic style ("Van Gogh"), (2) object ("Car"), and (3) celebrity ("Emma Watson"). The unlearned model exhibits a clear suppression in the target concept while retaining general image quality. The bottom row demonstrates generation with unrelated prompts, confirming that the model maintains its generative capacity outside the erased concept domain.

representation, enabling their targeted suppression without affecting unrelated outputs or requiring full retraining.

Since 2023, concept erasure has gained growing attention as a promising method for enhancing T2I generation safety and controllability. An increasing number of methods have been proposed to suppress sensitive concepts via architectural interventions, embedding manipulations, or plug-in modules. However, despite rapid method proliferation, the literature remains fragmented, lacking a unified framework for systematic comparison. This fragmentation results in unclear distinctions regarding the specific stages of model intervention, the granularity and completeness of erasure, and whether suppression occurs through prompt-level modification or output-level optimization, complicating method selection and practical adoption.

To date, only one prior survey [13] has explored concept erasure techniques, offering an initial yet broad classification. However, its categorization primarily emphasizes broader method types, such as finetuning, closed-form editing, and inference-time interventions, and has not yet included some recent methodological developments or detailed comparative analyses. In addition, it addresses practical deployment challenges, such as concept entanglement, semantic overlap, and incremental or multiconcept erasure, in a limited manner, leaving these increasingly important real-world issues insufficiently explored. In contrast, our survey introduces a comprehensive, multidimensional taxonomy, systematically comparing methods across intervention points and granularity, integrating advancements up to May 2025, and explicitly

addressing critical real-world deployment considerations. To further support practical understanding, we provide a qualitative evaluation of erasure effectiveness and model utility across different intervention levels and optimization strategies, offering actionable insights for selecting or designing appropriate methods for specific application scenarios.

To bridge these gaps, we present a structured and comprehensive survey of concept erasure techniques in text-to-image diffusion models. Our key contributions are as follows:

- **Comprehensive coverage of recent developments.** We compile and review a wide range of concept erasure methods proposed up to May 2025, including recent advances in plug-and-play modules, continual erasure strategies, and preference-aligned optimization, thus addressing the limited temporal scope in previous work.
- **Multi-dimensional taxonomy and framework construction.** As illustrated in Figure 2, we introduce a three-dimensional classification framework that organizes concept erasure techniques by optimization strategy, intervention level and semantic scope, thus addressing the lack of detailed technical categorization in prior work.
- **Evaluation protocols and benchmarking analysis.** We systematically summarize widely adopted evaluation metrics including Erasure Success Rate, FID, and CLIP Score, and analyze representative benchmark datasets designed for diverse downstream evaluation scenarios, thus facilitating reproducibility and rigorous comparative analyses.
- **Practical challenges and real-world constraints.** We

investigate challenges encountered in real-world applications, including conceptual combination erasure (CCE), erasure-utility trade-offs, and the evolving definitions of sensitive concepts, thus addressing the underexplored practical deployment issues in the literature.

Building on the motivations and contributions outlined above, the remainder of this paper provides a structured exploration of concept erasure techniques in T2I diffusion models. Section II reviews the architecture of text-to-image diffusion models and introduces key technical concepts relevant to concept erasure. Section III introduces a detailed, multi-dimensional taxonomy that systematically organizes concept erasure methods by intervention-level, optimization strategies and semantic scope, enabling clearer methodological comparisons and deeper insights into their respective strengths and limitations. Section IV provides an overview of the datasets, standard metrics, and commonly adopted benchmarks used in the assessment of concept erasure methods. Section V discusses practical challenges in deploying concept erasure methods in real-world systems and outlines open research questions and emerging directions, and Section VI concludes the paper.

II. BACKGROUND

This section introduces the technical foundations required for understanding concept erasure in T2I diffusion models, covering model architectures, cross-attention mechanisms, semantic concept definitions, and their distinctions from related tasks such as machine unlearning.

A. T2I Generation and Diffusion Models

Early approaches for T2I generation were based on GANs, but encountered significant limitations in scalability and training stability [1]–[3]. In contrast, diffusion models [7] have emerged as a dominant paradigm for high-fidelity image synthesis, leveraging a stochastic denoising process that gradually transforms noise into structured visual content. A diffusion model consists of two components: a forward process that incrementally corrupts data by adding Gaussian noise over T time steps, and a reverse process that learns to recover the data by denoising. Given an input image x_0 , the forward process is defined as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{(1 - \beta_t)}x_{t-1}, \beta_t I) \quad (1)$$

where β_t is a predefined variance schedule. The reverse process, often parameterized by a neural network ϵ_θ , predicts the noise component ϵ from a noisy sample x_t , enabling the model to reconstruct x_0 from a standard Gaussian prior $x_T \sim \mathcal{N}(0, I)$.

To improve computational efficiency, modern systems such as Stable Diffusion perform denoising in a lower-dimensional latent space. A variational autoencoder compresses images into latent representations z , and the denoising network ϵ_θ operates on these latents instead of full-resolution images. The conditioning information c is typically obtained from a frozen text encoder such as CLIP [98], whose outputs are injected into the denoising UNet [99] via cross-attention mechanisms.

B. Cross-Attention Mechanisms

Central to T2I diffusion models is the cross-attention mechanism, which dynamically aligns textual concepts with visual regions during generation, facilitating targeted semantic manipulations essential for concept erasure. Given a sequence of text embeddings $c \in \mathbb{R}^{L \times d}$, where L is the length of the sequence and d is the dimension of the embedding, and intermediate features of the image $f \in \mathbb{R}^{H \times W \times d'}$, the cross-attention module computes:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

where the query $Q = fW_Q$ is derived from the image features, and the key and value matrices $K = cW_K$, $V = cW_V$ are derived from the text embeddings. This mechanism enables each spatial location in the image to dynamically attend to relevant tokens, supporting fine-grained semantic alignment essential for concept erasure interventions.

C. Concepts and Sensitive Categories in T2I Diffusion Models

In the context of T2I diffusion models, a concept refers to a semantically meaningful and generative unit—ranging from concrete objects (e.g., “Eiffel Tower”), visual styles (e.g., “Van Gogh’s Starry Night”), and abstract notions (e.g., “freedom”), to socially sensitive or ethically charged categories (e.g., “violence”, “nudity”). Such concepts are implicitly embedded in the model’s latent space as a result of large-scale pretraining on multimodal data. A particularly critical subset of concepts is known as NSFW (Not Safe For Work) [91], [99], [100], which encompasses content considered inappropriate for general audiences due to its explicit, offensive, or harmful nature. This includes visual representations involving nudity, graphic violence, or hate symbols. The generation of NSFW content in T2I models raises significant ethical, legal, and societal concerns. Although service providers often deploy post-hoc content filters to suppress such outputs [24], [100], [101], these methods can be circumvented and lack semantic control. This underscores the necessity of concept erasure, to proactively and reliably remove unsafe content from the generative capabilities.

D. Concept Representation and Disentanglement Hypothesis

Concept erasure methods in T2I diffusion models are typically grounded on the assumption that semantic concepts are at least partially disentangled within the model’s internal representations [73]. This assumption of semantic disentanglement is fundamental to concept erasure, motivating targeted interventions within specific internal model structures. Empirically, conceptual information tends to exhibit spatial or channel-wise separability across architectural components: token embeddings within the text encoder often occupy distinct semantic subspaces; cross-attention maps align textual prompts with visual regions; and U-Net latent feature maps encode visual attributes along dedicated channel patterns. This architectural modularity has been exploited across various erasure strategies [30], [43], [61], including conditioning space modification at the embedding level, attention activation suppression at the cross-attention level, and feature masking within latent representations. This modular organization forms the structural basis for categorizing concept erasure methods according to their intervention targets, optimization strategies, and semantic scopes, as systematically analyzed in the following taxonomy.

E. Distinguishing Concept Erasure from Machine Unlearning

Concept erasure and machine unlearning [17]–[19] are both concerned with removing undesired information from models, but they fundamentally differ in objectives, intervention levels, and implementation strategies. Concept erasure operates at the semantic level, with

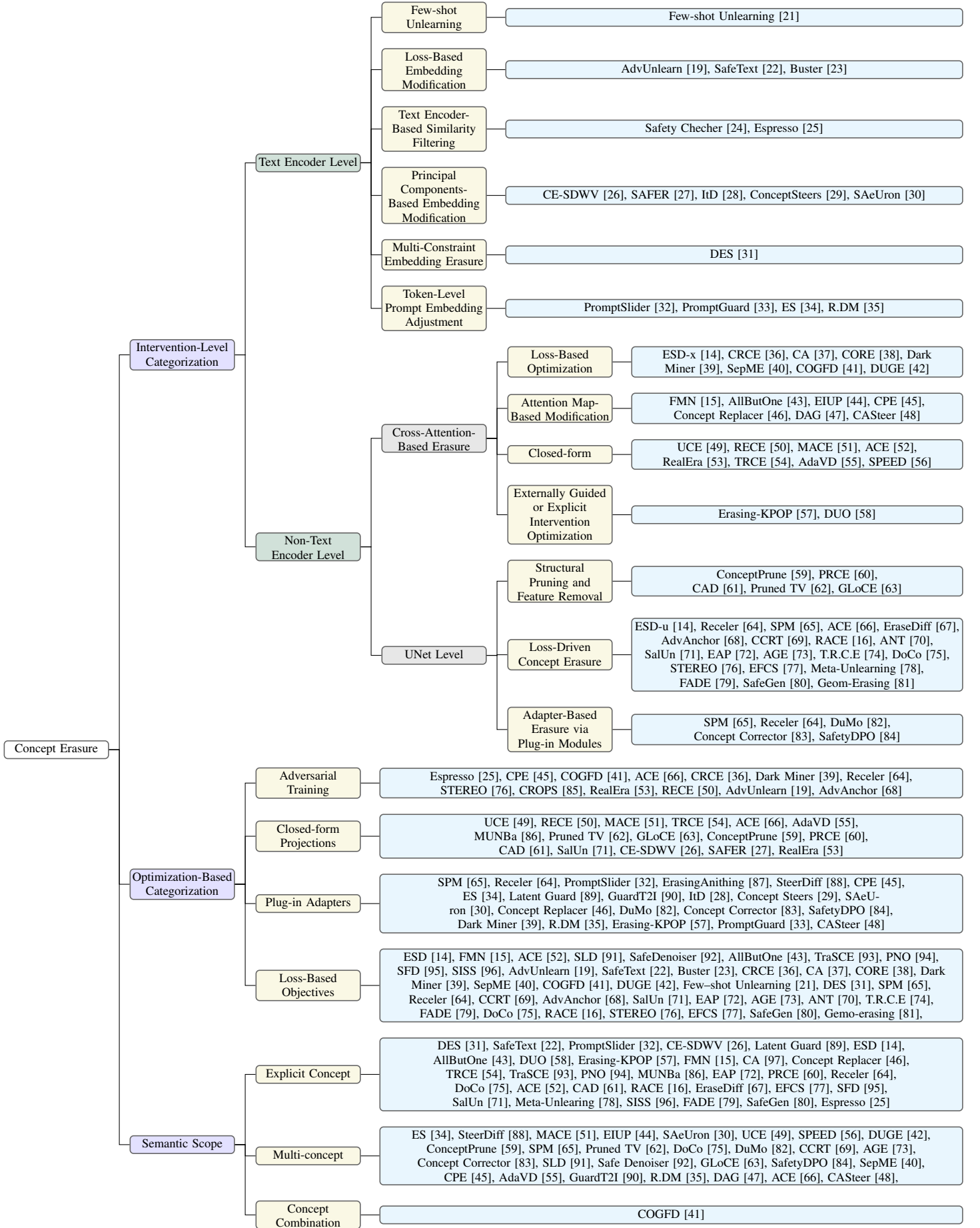


Fig. 2. Taxonomy of concept erasure techniques in text-to-image generative models, categorized along three orthogonal axes: (1) intervention level (e.g., text encoder vs. non-text encoder interventions), (2) optimization strategy (e.g., adversarial alignment, plug-in adapters, loss-based training), and (3) semantic scope (explicit vs. multi-concept). Representative methods are annotated with corresponding references.

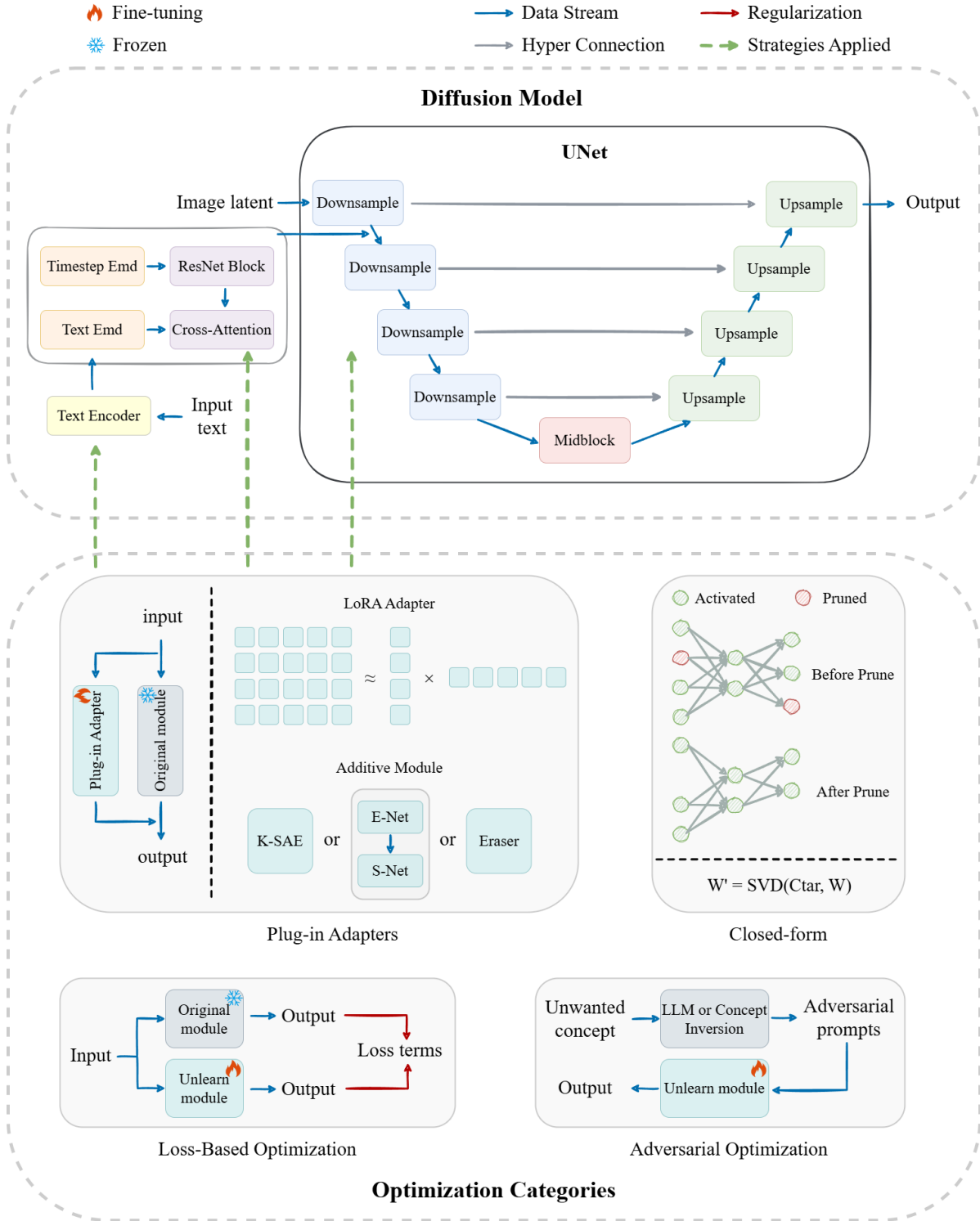


Fig. 3. Intervention-level categorization and corresponding optimization strategies for concept erasure in diffusion models. Concept erasure can be applied at different model components—text encoder, cross-attention, and UNet—each supporting specific methods such as loss-based optimization, plug-in adapters, or adversarial training.

the objective of suppressing a model’s ability to generate or express specific concepts, such as NSFW content, violent imagery, or artistic styles, without retraining the model on modified datasets. It assumes that semantic concepts are at least partially disentangled within internal representations, and typically intervenes through localized techniques such as gradient-based optimization [31], [40], [58], sub-

space projection [50], [51], [59], or plug-in modules [28], [64], [82]. Machine unlearning, in contrast, aims at retroactively eliminating the influence of particular training samples for privacy compliance (e.g., GDPR’s “right to be forgotten” [102]). It typically involves dataset scrubbing followed by retraining or fine-tuning, parameter pruning combined with re-optimization, or computationally intensive

influence function approximations. Such methods face scalability challenges in T2I models, where learned knowledge is typically distributed and not easily attributable to discrete training instances. Thus, concept erasure offers a complementary and practical approach for addressing semantic-level safety concerns.

III. TAXONOMY OF CONCEPT ERASING METHODS

As illustrated in Figure 2, we propose a structured taxonomy that systematically categorizes existing concept erasure methods for T2I diffusion models along three critical dimensions: the intervention level that identifies the specific architectural component targeted for erasure, the optimization strategy that refers to the underlying algorithmic mechanisms facilitating concept suppression, and the semantic scope defining the complexity and nature of the concepts being addressed. Compared to previous work, which typically provides broader categorizations, our multi-dimensional framework offers a more detailed and nuanced characterization of each method. This structure allows for clearer methodological comparisons and provides deeper insights into each approach’s underlying rationale and practical applicability. In the subsequent subsections, we rigorously analyze each dimension, providing a comprehensive and systematic perspective that enhances understanding beyond existing categorizations.

A. Intervention-Level Categorization

The architecture of text-to-image diffusion models presents multiple levels of semantic intervention, where concept erasure techniques can operate with varying granularity and control. These interventions primarily fall into two overarching categories: Text Encoder Level and Non-Text Encoder Level. Text Encoder-level approaches operate at the earliest stage, modifying prompt embeddings to suppress target concepts before semantic propagation begins. Non-Text Encoder-level approaches operate deeper within the generative architecture, further subdivided into two distinct levels: (1) Cross-Attention-Level, which intervenes at the alignment stage between text and image, typically by manipulating attention maps or key/value projections; and (2) UNet-Level, targeting the internal feature dynamics within the UNet backbone, reshaping the denoising process through plug-in modules, feature masking, or subspace projection. The following subsections provide a detailed analysis of methods according to this hierarchical categorization.

1) Text Encoder Level

The text encoder in most T2I diffusion models—typically based on CLIP [98]—serves as a unified semantic interface between language and image domains. This architectural consistency enables efficient and transferable concept erasure strategies. By modifying the prompt embedding space, specific concepts can be decoupled from their downstream visual influence while maintaining overall model utility. Compared to full-model finetuning, such interventions are lightweight [26], [28], [29] and often compatible across models. Figure 3 illustrates typical intervention sites within the text encoder pipeline. Specifically, (a) modifies the original text encoder via fine-tuning, (b) introduces additional modules into the encoder to steer its representations, and (c) intervenes at the interface between the text encoder and the downstream UNet.

Few-shot Unlearning Few-shot Unlearning [21] proposes a lightweight concept erasure strategy that operates under limited data availability, aiming to suppress target semantics with minimal retraining overhead. It updates the MLP and final self-attention layers using only 4-shot real images per target concept. The optimization minimizes the diffusion reconstruction loss:

$$\underset{\theta}{\text{minimize}} \mathbb{E}_{x, \epsilon, t, y} [\|\epsilon - \epsilon_{\theta}(x, t, c_{\theta}(y))\|_2^2], \quad (3)$$

where ϵ_{θ} is the denoising network and $c_{\theta}(y)$ denotes the learned text embedding. Although effective for well-localized concepts, this

method struggles with abstract or highly entangled categories [21] such as “church” or “freedom”.

a) Loss-Based Embedding Modification

Several methods implement concept erasure by modifying the training loss applied to text encoder embeddings. AdvUnlearn [19] formulates erasure as a bilevel optimization problem. It introduces a regularization term on retained concepts to balance the erasure objective and maintain overall utility. The total loss is based on a combination of $\mathcal{L}_{ESD}$ and:

$$\mathcal{L}_{\text{retain}} = \mathbb{E}_{c \in C_{\text{retain}}} [\|\epsilon'_{\theta}(x_t, c) - \epsilon_{\theta}(x_t, c)\|_2^2], \quad (4)$$

where ϵ_{θ} is the denoising model and c denotes prompts associated with non-target (retained) concepts.

SafeText [22] introduces dual objectives: one to pull safe prompt embeddings closer to the original representation, and one to push unsafe prompts away from their prior representations. The two losses are defined as:

$$\mathcal{L}_{\text{retain}} = -\mathbb{E}_{P_{un} \sim \mathbb{D}_{un}} [d_e(\mathcal{T}_s(P_{un}), \mathcal{T}(P_{un}))], \quad (5)$$

$$\mathcal{L}_{\text{erase}} = \mathbb{E}_{P_s \sim \mathbb{D}_s} [d_u(\mathcal{T}_s(P_s), \mathcal{T}(P_s))], \quad (6)$$

where d_e is Euclidean distance and d_u is negative absolute cosine similarity. $\mathbb{D}_s$ and $\mathbb{D}_{un}$ denote unsafe and safe prompt datasets, respectively.

Buster [23] addresses concept erasure via adversarial backdoor suppression:

$$y^{(n+1)} \leftarrow y^{(n)} - \eta \nabla E(y^{(n)}) + \epsilon^{(n)} \quad (7)$$

where y is the original input, $\eta > 0$ is learning rate, $\epsilon^{(n)} \in \mathcal{N}(0, \sigma)$ is the random Gaussian noise sampled at step n . It constructs two sets of losses: one that encourages consistency over benign prompts, and another that penalizes misalignment between benign and adversarially triggered representations:

$$\mathcal{L}_{\text{Benign}} = \frac{1}{|B|} \sum_{c \in B} d(\mathcal{T}(c), \tilde{\mathcal{T}}(c)), \quad (8)$$

$$\mathcal{L}_{\text{Backdoor}} = \frac{1}{|B'|} \sum_{w \in B'} d(\mathcal{T}(t), \tilde{\mathcal{T}}(w)), \quad (9)$$

where t is the target concept, w is a backdoor variant, and d measures representation distance (e.g., cosine or Euclidean).

b) Text Encoder-Based Similarity Filtering

Unlike model-level interventions that modify internal representations, filter-based methods operate at the output level to detect and block undesirable content. These approaches monitor the semantic similarity between generated outputs and predefined concept sets, triggering suppression when similarity exceeds a threshold. The Safety Checker [24] is a widely adopted example. It computes the cosine similarity between a generated image and concept embeddings (e.g., nudity or violence), and blocks the image when the similarity exceeds a predefined threshold. Building on this idea, Espresso [25] proposes a similarity-based detection-and-suppression mechanism. It compares the embedding of the generated image x with two concept sets: undesired (c^u) and acceptable (c^a). A filtering function F evaluates the relative closeness of x to each set, and suppression is triggered when proximity to unsafe concepts exceeds a threshold Γ :

$$F_{SD}(x) = \begin{cases} 1 & \text{if } F(x, c^u, c^a) > \Gamma, \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

The function $F(x, c^u, c^a)$ computes a margin-based similarity between the generated image x and two sets of concepts: undesired (c^u) and acceptable (c^a). If the score exceeds a predefined threshold Γ , the image is flagged and suppressed. To improve resilience against prompt-based attacks, Espresso fine-tunes the text encoder using adversarial and synthetic examples to sharpen semantic boundaries. While not modifying the model’s internal representations, Espresso offers a lightweight alternative for post-hoc content filtering.

c) Principal Components-Based Embedding Modification

This class of methods performs concept erasure by identifying concept-specific subspaces within the text embedding space. By estimating principal components associated with a target concept, using techniques such as singular value decomposition (SVD) or sparse autoencoders (SAE), these approaches suppress the influence of the concept through gradient projection, subspace masking, or component suppression.

SVD-based erasure methods estimate a concept subspace by performing singular value decomposition on a set of concept-aligned prompt embeddings. The learned subspace serves as a geometric prior for suppressing undesired concepts while retaining general model capability. CE-SDWV [26] constructs the concept subspace S_c by applying SVD to the covariance matrix Σ_{ec} of embeddings associated with concept c :

$$U, \Sigma, V = \text{SVD}(\Sigma_{ec}). \quad (11)$$

The first left singular vector U_1 is selected as the principal direction for S_c , and gradient updates to the text encoder are orthogonalized to this direction to avoid reintroducing erased concepts. SAFER [27] improves this approach in two ways. It first applies text inversion to optimize the embedding distribution used for subspace construction. Then, it expands the subspace by aggregating multiple semantically related concepts (e.g., “Van Gogh”, “Starry Night”, and “Sunflower”) to enable broader erasure of stylistic concepts. This strategy enhances both the specificity and generalization capacity of subspace-guided suppression.

Sparse autoencoder-based approaches learn a disentangled basis over prompt embeddings by encoding semantic concepts as sparse component activations. These methods identify the most salient components associated with the target concept and suppress them selectively during inference or conditioning. ItD [28], ConceptSteers [29], and SAeUron [30] integrate sparse autoencoders into the text encoder pipeline. Given an input embedding $\mathbf{e}$, the encoder identifies Top- k concept-relevant components $\mathbf{z}$ via a linear projection, and reconstructs a cleaned prompt embedding $\hat{\mathbf{e}}$ using only these components:

$$\begin{aligned} \mathbf{z} &= \text{TopK}(W_{\text{enc}}(\mathbf{e} - \mathbf{b}_{\text{pre}})), \\ \hat{\mathbf{e}} &= W_{\text{dec}}\mathbf{z} + \mathbf{b}_{\text{pre}}, \end{aligned} \quad (12)$$

where W_{enc} , W_{dec} denote the encoder and decoder weights, and $\mathbf{b}_{\text{pre}}$ is a learned bias term. The resulting representation $\hat{\mathbf{e}}$ excludes components attributed to the erased concept, thereby suppressing it while retaining unrelated semantics. Compared to SVD-based methods, SAE-based models can flexibly learn task-specific representations and adapt to concept co-occurrence, albeit with less interpretability.

Principal component-based methods offer interpretable and modular strategies for disentangling semantic concepts in the embedding space. However, their effectiveness depends heavily on the accuracy of subspace estimation and the structural regularity of the underlying concept distribution.

d) Multi-Constraint Embedding Erasure

DES [31] refines concept erasure at the text encoder level by introducing a structured three-term objective (Eq. 13) designed to simultaneously neutralize unsafe prompts, preserve benign semantics, and suppress domain-specific content such as nudity. Specifically, the method defines: (1) Unsafe Embedding Neutralization (UEN), which encourages the embedding of unsafe prompts to align with a neutral subspace; (2) Safe Embedding Preservation (SEP), which stabilizes the representations of benign prompts under perturbation; and (3) Nudity Embedding Neutralization (NEN), which explicitly suppresses sexually explicit content by discouraging alignment with unconditioned guidance vectors. The corresponding loss terms are

formulated as follows:

$$\begin{aligned} \mathcal{L}_{\text{UEN}} &= \frac{1}{B} \sum_{i=1}^B \left(1 - \frac{\tilde{\mathbf{e}}_{u,i} \cdot \hat{\mathbf{e}}_{s,i}}{\|\tilde{\mathbf{e}}_{u,i}\| \|\hat{\mathbf{e}}_{s,i}\|} \right), \\ \mathcal{L}_{\text{SEP}} &= \frac{1}{B} \sum_{i=1}^B \left[\left(1 - \frac{\tilde{\mathbf{e}}_{s,i} \cdot \mathbf{e}_{s,i}}{\|\tilde{\mathbf{e}}_{s,i}\| \|\mathbf{e}_{s,i}\|} \right) + \left(1 - \frac{\tilde{\mathbf{e}}'_{s,i} \cdot \mathbf{e}_{s,i}}{\|\tilde{\mathbf{e}}'_{s,i}\| \|\mathbf{e}_{s,i}\|} \right) \right], \\ \mathcal{L}_{\text{NEN}} &= 1 - \frac{\tilde{\mathbf{e}}_n \cdot \mathbf{e}_{uc}}{\|\tilde{\mathbf{e}}_n\| \|\mathbf{e}_{uc}\|}. \end{aligned} \quad (13)$$

Together, these three loss components constitute a structured and concept-specific regularization framework, enabling DES to enforce targeted suppression of unsafe concepts while preserving the integrity of benign semantic representations within the text encoder space.

e) Token-Level Prompt Embedding Adjustment

This class of methods performs concept erasure by modifying the prompt embeddings at the token level without altering model parameters. These techniques fall into two major categories: (1) introducing learnable auxiliary embeddings to semantically offset the original prompt, and (2) dynamically modulating token embeddings based on their estimated risk or importance. Both approaches offer lightweight, model-agnostic means of adjusting semantic conditioning.

PromptSlider [32] exemplifies the first category. It introduces a learnable auxiliary prompt vector S^* that semantically offsets the original prompt in the embedding space. The optimization objective minimizes denoising error during generation:

$$S^* = \arg \min_S \mathbb{E}_{z,y,\epsilon,t} \|\epsilon_t - \epsilon_\theta(z_t, p_\theta(y, S), t)\|_2^2, \quad (14)$$

where z_t is the latent at timestep t , $p_\theta(y, S)$ is the composed embedding of the original prompt y and slider S , and ϵ_θ is the denoising model. PromptGuard [33] builds upon this idea by learning a soft suffix embedding P^* , appended to the prompt. A contrastive loss is used to push concept-bearing prompts farther from unsafe representations, while pulling them closer to neutral concepts, thereby enhancing alignment robustness.

ES [34] illustrates the second category, using token-aware risk scores to suppress harmful tokens in the embedding space. The adjusted embedding is computed as:

$$\text{Emb}_c = \text{Emb}_s - \alpha \cdot S \cdot \text{Emb}_t, \quad (15)$$

where Emb_s is the original embedding, Emb_t is the toxic token vector, and S is a learned risk score. The scalar α controls suppression intensity.

RDM [35] leverages LoRA to learn the semantic direction S of a given concept c . By adding or subtracting S from the original prompt embedding, the method can selectively enhance or suppress the visual manifestation of the concept in the generated image.

f) Summary of Text Encoder-Level Concept Erasure Methods

Text encoder-level concept erasure methods offer lightweight and principled intervention mechanisms but exhibit diverse trade-offs between precision, efficiency, and semantic coverage. Few-shot unlearning and loss-based embedding optimization enable effective semantic suppression with minimal architectural changes. Subspace-based and token-level adjustment strategies provide finer-grained semantic control, though often at the cost of increased complexity or reduced robustness. Similarity filtering offers model-agnostic deployment flexibility but remains limited to surface-level prompt representations. While text encoder-level approaches provide efficient manipulation, their reliance on prompt embeddings constrains their capacity for spatially precise and semantically robust concept control. This motivates direct intervention at the cross-attention layer, where textual and visual features are dynamically aligned.

2) Non-Text Encoder Level

While text encoder-level methods intervene at the source of semantic injection, non-text encoder approaches operate deeper within

the generative architecture, targeting latent representations along the model's internal pathways. These include the cross-attention layer, which governs text-to-image alignment, and the UNet backbone, which drives iterative feature refinement during image synthesis. By operating on internal activations—such as attention maps [15] or latent feature channels [59]—these methods enable concept suppression without modifying the prompt itself. In what follows, we first examine methods that intervene at the cross-attention layer to disrupt semantic alignment, followed by approaches that operate within the UNet to reshape the generative trajectory.

a) Cross-Attention-Based Erasure

In text-to-image diffusion models [7]–[9], the cross-attention layer plays a pivotal role in aligning textual semantics with spatial features during the denoising process. It governs how each token in the prompt contributes to specific regions in the generated image via learned attention maps. As such, cross-attention provides a semantically interpretable and structurally modular entry point for concept erasure. By intervening in attention computation, either through attention weight suppression or modification of key/value projections, methods can selectively suppress the influence of undesired concepts while preserving unrelated content.

Loss-Based Optimization Loss-based erasure methods at the cross-attention level formulate concept suppression as a targeted optimization objective. By adjusting loss functions associated with attention mechanisms or conditional noise prediction, these approaches weaken the influence of undesired concepts without modifying the underlying network architecture. Representative methods differ in whether they align conditional scores, utilize anchor-based comparisons, or apply structured multi-concept regularization.

ESD-x [14] suppresses concept activation by minimizing the discrepancy between conditional and unconditional noise predictions. The erasure objective is formulated as an ℓ_2 loss:

$$\mathcal{L}_{\text{ESD}} = \|\epsilon_\theta(x_t, c, t) - \epsilon_{\theta^*}(x_t, c, t)\|_2^2, \quad (16)$$

where c denotes the target concept and ϵ_{θ^*} represents the prediction of a concept-neutral reference model. CRCE [36] extends the ESD-x framework by incorporating auxiliary constraints on adversarial coreferent concepts and benign concepts to be retained. The full loss function is:

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\text{ESD}} + \frac{1}{M} \sum_{c' \in \text{Coref}} \mathcal{C}_{c'} \|\epsilon_\theta(x_t, c', t) - \epsilon_{\theta^*}(x_t, c, t)\|_2^2 + \\ \frac{1}{N} \sum_{c'' \in \text{Retain}} \mathcal{C}_{c''} \|\epsilon_\theta(x_t, c'', t) - \epsilon_{\theta^*}(x_t, c'', t)\|_2^2, \end{aligned} \quad (17)$$

where M and N are the number of coreferent and retained concepts, respectively.

To address semantic overlap between similar prompts, CA [37] introduces the notion of anchor concepts and defines contrastive losses that align model behavior between target and neutral prompts. It proposes two complementary objectives. The first is a KL divergence loss over diffusion features:

$$\mathcal{L}_{\text{model}}(\mathbf{x}, \mathbf{c}, \mathbf{c}^*) = \mathbb{E} \left[w_t \left\| \hat{\Phi}(\mathbf{x}_t, \mathbf{c}, t) - \hat{\Phi}(\mathbf{x}_t, \mathbf{c}^*, t) \right\| \right], \quad (18)$$

and the second is a noise-based loss applied to anchor prompts:

$$\mathcal{L}_{\text{noise}}(\mathbf{x}, \mathbf{c}, \mathbf{c}^*) = \mathbb{E} \left[w_t \left\| \epsilon - \hat{\Phi}(\mathbf{x}_t, \mathbf{c}^*, t) \right\| \right]. \quad (19)$$

CORE [38] improves upon CA [37] by introducing adaptive proxy prompts, which are dynamically selected rather than fixed placeholders. The alignment and retention objectives are:

$$\begin{aligned} \mathcal{L}_f = \mathbb{E} \left[\|\epsilon_\theta(x_t, p_f) - \epsilon_\theta^*(x_t, p_a)\|_2^2 \right], \\ \mathcal{L}_r = \mathbb{E} \left[\|\epsilon_\theta(x_t, p_r) - \epsilon_\theta^*(x_t, p_r)\|_2^2 \right], \end{aligned} \quad (20)$$

where p_f , p_r , and p_a are prompts from the forget, retain, and anchor sets, respectively.

As concept entanglement becomes increasingly common in T2I generation, several methods extend loss-based optimization to disentangle multiple concepts simultaneously. To achieve this goal, Dark Miner [39] initiates a three-stage pipeline involving concept mining, verification, and selective gradient suppression. Its erasure and retention objectives are jointly optimized:

$$\begin{aligned} l_{\text{erase}} &= \mathbb{E}_{x,t,\epsilon} [\|\epsilon_\theta(x_t|t, c) - \epsilon_{\theta_0}(x_t|t, c_0)\|_2^2] \\ l_{\text{retain}} &= \mathbb{E}_{x,t,\epsilon} [\|\epsilon_\theta(x_t|t, c_0) - \epsilon_{\theta_0}(x_t|t, c_0)\|_2^2] + \\ &\quad \mathbb{E}_{x,t,\epsilon} [\|\epsilon_\theta(x_t|t, c) - \epsilon_{\theta_0}(x_t|t, 0_c)\|_2^2] + \\ &\quad \mathbb{E}_{x,t,\epsilon} [\|\epsilon_\theta(x_t|t, -c) - \epsilon_{\theta_0}(x_t|t, -c)\|_2^2] \end{aligned} \quad (21)$$

SepME [40] formulates multi-concept disentanglement as temporally constrained optimization combined with weight decomposition, restricting parameter updates to concept-specific subspaces:

$$\begin{aligned} \mathcal{L}_{\text{SepME}} &= \sum_{i=1}^N \eta_i \mathcal{L}_{\text{cor}}(c_{i,f}, \Delta\theta_{i,d_m}) + \lambda \|W\|_p, \\ \text{s.t. } \forall i \in [1, N], \mathcal{L}_{\text{cor}}(c_{i,f}, \Delta\theta_{i,d_m}) &= 0, \end{aligned} \quad (22)$$

where W are attention parameters and $\Delta\theta$ are learned weight increments.

Beyond direct disentanglement at the parameter level, recent methods seek to model higher-order semantic relationships or integrate multiple regularization objectives to enhance robustness in multi-concept erasure. In this context, COGFD [41] and DUGE [42] propose novel frameworks based on logic-driven concept graphs and unified multi-objective optimization, respectively. COGFD [41] introduces a logic-aware concept disentanglement mechanism by constructing a concept graph via LLM agents. Erasure is then guided by a combination of gradient ascent on the main target and gradient descent on related components:

$$\mathcal{L} = \alpha \cdot \exp(-D(\phi_\theta, \phi'_\theta, \hat{m})) - (1 - \alpha) \cdot \exp\left(\sum_{i=1}^k D(\phi_\theta, \phi'_\theta, \hat{c}_i)\right), \quad (23)$$

where ϕ'_θ is the frozen reference model and $\hat{m}, \{\hat{c}_i\}$ are concept group embeddings.

DUGE [42] unifies attention suppression, retention, and regularization into a multi-objective erasure framework:

$$\begin{aligned} \mathcal{L}_U &= \frac{1}{M} \sum_{i=1}^M (A_{\theta_c}^i - A_{\phi_n}^i)^2, \\ \mathcal{L}_{pr} &= \mathbb{E} \left[\left\| \epsilon - \epsilon_\theta^\delta(x_t, c_{pr}, t) \right\|_2^2 \right], \\ \mathcal{R}_{pr} &= \min_{\theta^\delta} \mathcal{D}_{\text{KL}} \left(\theta^\delta \parallel \theta^{\delta-1} \right), \end{aligned} \quad (24)$$

where A_{ϕ_n} represents the negative attention map, and c_{pr} is a safe concept.

Attention Map-Based Modification Cross-attention maps [97] in T2I models serve as a crucial interface that dynamically aligns textual concepts with spatial regions during generation. Manipulating these maps offers a fine-grained and interpretable pathway for concept suppression. By directly modulating attention activations, methods can selectively disrupt or redirect the model's dependency on undesired features while preserving general semantics.

Early approaches [15], [43] focus on direct suppression of attention activations associated with target concepts. FMN [15] introduces an Attention Resteering mechanism, which identifies context embeddings linked to target concepts, computes their attention maps, and minimizes these activations during training:

$$\mathcal{L}_{\text{FMN}} = \sum_{a_t \in A_t} \left\| a_t^{[n]} \right\|^2, \quad (25)$$

where A_t denotes the set of attention maps associated with the target concept across heads and layers. AllButOne [43] also targets

direct suppression by manipulating activation patterns within the cross-attention layers. It blocks spatial localization by attenuating the attention weights corresponding to the target concept, balancing erasure strength against preservation of unrelated features.

Beyond direct suppression, several methods [44] augment or reweight attention maps to allow adjustable control over erasure strength. EIUP [44] concatenates the original attention map M_t with a weighted version of the target concept attention map M_t^* . A sequence of learnable weights W controls the degree of influence from the target prompt:

$$\begin{aligned}\hat{M}_t &= \text{Concat}(M_t, WM_t^*), \\ \hat{V}_t &= \text{Concat}(V_t, V_t^*), \\ \hat{A}_t &= \hat{W}_t \times \hat{V}_t,\end{aligned}\quad (26)$$

where M_t , M_t^* , and $\hat{M}_t$ are the original, target-specific, and augmented attention maps, respectively. CPE [45] applies adversarial training [103] combined with lightweight residual attention gates (ResAGs) to perturb attention responses tied to sensitive concepts, enhancing model resilience while maintaining generation quality.

Building upon attention feature modulation, several methods [46]–[48] further refine concept control by steering or replacing latent semantic activations within cross-attention layers. These approaches differ in how they identify, manipulate, and regulate concept-specific attention patterns during generation. Concept Replacer [46] performs discrete latent concept replacement during denoising. A dedicated concept localizer detects features associated with the target concept, which are then redirected toward semantically neutral replacements via a Dual Prompt Cross-Attention (DPCA) module. DAG [47] extends latent attention manipulation by introducing a two-stage optimization framework. Guideline Detection refines token embeddings using annotated datasets to improve the accuracy of generated cross-attention maps, while Safe Self-Regulation dynamically adjusts diffusion guidance strength based on confidence scores derived from the CAMs. CASTeering [48] advances continuous control by learning steering vectors from positive and negative concept examples. For a given target concept, the steering vector is constructed by averaging activation differences across examples. The learned vectors are then dynamically applied to modify attention activations, enabling fine-grained enhancement or suppression of target concepts during generation.

Closed-form Closed-form optimization provides efficient and interpretable interventions for concept erasure in T2I diffusion models. These methods bypass iterative gradient-based updates by applying analytical projections, typically modifying the key (K) and value (V) matrices in cross-attention layers. This line of work offers fast execution, low memory overhead, and stable erasure dynamics, making them well-suited for multi-concept or real-time scenarios.

UCE [49] is the pioneering closed-form approach, which suppresses target concepts by analytically updating the K and V projection weights associated with concept-aligned tokens. Specifically, it computes a modified projection matrix W using a closed-form least-squares update that preserves safe concepts and neutralizes target concepts:

$$W = \left(\sum_{c_i \in E} W^{old} c_i^* c_i^T + \sum_{c_j \in P} W^{old} c_j c_j^T \right) \cdot \left(\sum_{c_i \in E} c_i c_i^T + \sum_{c_j \in P} c_j c_j^T \right)^{-1} \quad (27)$$

where E and P denote the sets of target and preserved concepts, c_i and c_j are their respective embeddings, and c_i^* is the surrogate concept embedding used to replace c_i .

Extending the foundation laid by UCE [49], subsequent methods introduce structured enhancements, ranging from adversarial robustness to scalable multiconcept suppression and semantically guided regulation. RECE [50] incorporates adversarial prompt mining and

regularized projection, achieving efficient erasure in under 3 seconds per concept while maintaining generation fidelity. MACE [51] combines LoRA-based low-rank adaptation with closed-form projection to support scalable multi-concept erasure (up to 100 concepts), balancing generalization and specificity. ACE [66] formulates attention gating via concept-specific gating vectors g , enabling interpretable masking at the token level, and integrates adversarial training to enhance robustness. RealEra [53] extends closed-form projection with neighborhood concept mining and beyond-concept regulation to fine-tune the UNet’s behavior. Its loss includes adversarial and preservation-aware terms:

$$\begin{aligned}\mathcal{L}_{\text{noise}} &= \|\epsilon_\theta(z_t, t, p'_c) - \hat{\epsilon}_\theta(z_t, t, p_{c^*})\|_2^2 \\ &\quad + \|\epsilon_\theta(z_t, t, p''_c) - \hat{\epsilon}_\theta(z_t, t, p''_c)\|_2^2,\end{aligned}\quad (28)$$

where p'_c is an adversarial prompt, p''_c a perturbed benign prompt, and $\epsilon_\theta/\hat{\epsilon}_\theta$ denote unlearned and original noise predictors.

Several methods extend closed-form erasure by incorporating subspace projection [56], trajectory adjustment [54], or geometric regularization [55] to enhance semantic precision. TRCE [54] builds on closed-form erasure by introducing a two-stage refinement: it first nullifies adversarially mined prompts via projection, then adjusts the denoising trajectory through contrastive optimization. AdaVD [55] performs spatial concept projection by orthogonally projecting prompt embeddings away from target token directions in the attention value space, using a cosine similarity metric $\delta(\cdot, \cdot)$ to minimize utility impact. SPEED [56] applies SVD-based subspace decomposition to compute a projection matrix S_c , constraining model updates Δ to the null space of preserved semantics and enabling precise suppression of target concepts.

Externally Guided or Explicit Intervention Optimization Beyond internal projection and suppression strategies, a different class of concept erasure methods employs externally guided objectives, such as prompt auxiliary injection and optimization aligned with preferences, to exert explicit control over semantic suppression [57], [58]. These externally guided interventions enable precise control over the erasure behavior, aligning the model with auxiliary embeddings or user-defined preferences.

Erasing-KPOP [57] injects learnable auxiliary prompts S^* into the cross-attention layers. These prompts are trained to absorb the semantics of the target concept, thereby displacing its influence from the latent space and weakening its generative impact. DUO [58] implements preference-aligned erasure using a two-stage process. Controlled Pairwise Data Generation first synthesizes matched image pairs (x_0^+, x_0^-) via SDEdit [104], with x_0^+ being the preferred (safe) image and x_0^- the unsafe counterpart. Direct Preference Optimization (DPO) is then applied to tune the model’s behavior, encouraging generation of x_0^+ over x_0^- when prompted with c_{tar} . While their dependence on labeled data or preference modeling may pose scalability challenges, these methods nonetheless demonstrate the potential of externally guided objectives to deliver controllable, interpretable, and semantically precise concept erasure.

b) UNet-Level Methods

Beyond semantic alignment, concept erasure can also intervene in the generative backbone of diffusion models. The UNet serves as the core component responsible for denoising and progressively refining latent features into realistic images. Methods at this level suppress concept expression by manipulating internal activations—through feature masking, structural pruning, or gradient-guided updates—within the denoising pathway. Compared to cross-attention interventions, UNet-level methods achieve deeper semantic decoupling by intervening directly in the visual feature synthesis progress [70], [79].

Structural Pruning and Feature Removal Several UNet-level methods adopt closed-form formulations analogous to UCE, targeting internal components such as neurons or subspaces for efficient concept suppression. ConceptPrune [59] and PRCE [60] suppress target concepts by identifying and pruning neurons whose activations are strongly correlated with concept-specific features. While this

achieves strong suppression, it may degrade model utility due to coarse granularity at the neuron level [59]. CAD [61] defines a concept importance objective $J(c, w)$, measuring how much a model component w contributes to the generation of concept c :

$$J(c, w) = \mathbb{E}_{(x,t)} [\|f(x_t, c) - f_{-w}(x_t, c)\|^2], \quad (29)$$

where f_{-w} denotes the model with component w masked. High-scoring components are pruned to erase the target concept. Pruned TV [62] extracts a task vector τ by fine-tuning on target-only prompts and subtracts it from the original model parameters:

$$\tau = \theta - \theta^{(c)}, \quad \hat{\theta} = \theta - \tau, \quad (30)$$

where θ is the original model, $\theta^{(c)}$ is trained on concept c , and $\hat{\theta}$ is the erased model. GLoCE [63] performs closed-form concept projection using LoRA, and suppresses principal components via a gated mechanism. Its output function is defined as:

$$f(X) = (1 - s(X))X + s(X)(P^*X + b^*), \quad (31)$$

$$s(X) = \sigma \left(\alpha^* \left(\left\| (V^*)^\top (X - \beta^*) \right\|_2^2 - \gamma^* \right) \right),$$

where $f(\cdot)$ is the T2I model, $s(X)$ denotes the gating function, and (P^*, b^*) are closed-form parameters projecting X away from concept-aligned components.

These closed-form strategies offer fast and modular concept suppression at the structural level. However, they rely on strong assumptions about concept localization and may compromise generalization if pruning or projection is overly aggressive [59], [61].

Loss-Driven Concept Erasure Loss-based optimization methods at the UNet level aim to suppress concept representations by introducing targeted objectives or regularization terms into the training process. These approaches typically operate on the noise prediction model ϵ_θ , modifying how it responds to prompts containing the target concept. We categorize these methods into five categories based on their loss structure and optimization goals.

A prominent subclass of UNet-level loss-based methods builds directly on the $\mathcal{L}_{\text{ESD}}$ formulation, adapting it to internal UNet activations. These include extensions such as ESD-u, Receler, SPM, and CCRT, minimizing the discrepancy between conditional and unconditional noise predictions to suppress concept activation. ESD-u extends this to the UNet by applying the loss:

$$\mathcal{L}_{\text{ESD}} = \|\epsilon_\theta(x_t, c, t) - \epsilon_\theta^*(x_t, t)\|^2, \quad (32)$$

where c is the target concept and ϵ_θ^* denotes the unconditioned noise prediction. Receler [64] and SPM [65] enhance $\mathcal{L}_{\text{ESD}}$ by incorporating LoRA-based fine-tuning into UNet, enabling modular and data-efficient suppression. AdvAnchor [68] introduces dynamic anchor concepts to better balance erasure effectiveness and utility preservation. CCRT [69] adds an MSE-based regularization term $\mathcal{L}_{\text{reg}}$ over safe prompts and leverages prompt calibration through LLMs and genetic algorithms.

Another line of work emphasizes utility preservation by introducing multi-objective loss functions. These methods, including SalUn [71], EAP [72], AGE [73], ANT [70], T.R.C.E [74], FADE [79], and DoCo [75], jointly balance concept suppression with non-target feature retention, often via regularization, auxiliary optimization, or domain alignment. SalUn [71] identifies salient weight regions and introduces an MSE regularization term over retained data to protect model generalization:

$$\mathcal{L}_{\text{SalUn}} = \mathcal{L}_{\text{model}}(x, c, c^*) + \beta \cdot \mathcal{L}_{\text{MSE}}(\theta_u, D_r), \quad (33)$$

where θ_u denotes the updated model parameters and D_r is the remaining dataset. EAP [72] proposes a bilevel optimization framework where adversarial preserving concepts c_a are selected from a Gumbel-Softmax distribution π over the concept set $R = C \setminus E$. The objective stabilizes erasure while preserving sensitive non-target concepts:

$$\min_{\theta'} \max_{\pi \in \Delta_R} \mathbb{E}_{c \in E} [\|\epsilon_{\theta'}(c) - \epsilon_\theta(c_n)\|^2] \quad (34)$$

$$+ \lambda \cdot \|\epsilon_{\theta'}(G(\pi) \odot R) - \epsilon_\theta(G(\pi) \odot R)\|^2,$$

where $G(\pi)$ denotes the relaxed sampling of concept embeddings via the temperature-controlled Gumbel-Softmax operator. AGE [73] generalizes this idea by continuously optimizing over semantically similar but non-identical concepts, replacing discrete override prompts with a soft distributional search strategy. ANT [70] focuses on mid-to-late denoising stages and applies a dual loss formulation— $\mathcal{L}_f$ to enforce forgetting and $\mathcal{L}_r$ to preserve relevant features—thus preventing late-stage reactivation of target concepts. T.R.C.E [74] improves distributional alignment by replacing the standard ℓ_2 regularization with the Cramér-Wold distance and the Jensen-Shannon divergence, enhancing stability under complex erasure trajectories. FADE [79] defines three jointly optimized loss terms—erasure, adjacency preservation, and guidance—based on the adjacency set $\mathcal{A}(c_{\text{tar}})$, ensuring target suppression while maintaining conceptual continuity. DoCo [75] aligns anchor and target concept output distributions via adversarial training and retains important gradient directions for benign concepts using a projection-based gradient filtering mechanism. These multi-objective losses reflect a shift from pure suppression to more nuanced balance between erasure strength and representational utility, particularly under distribution shift or multi-concept overlap.

Beyond direct loss engineering, some methods adopt adversarial or meta-optimization strategies to refine concept erasure dynamics. RACE [16] enhances robustness against prompt perturbations through adversarial training on concept-conditioned noise predictions. STEREO [76] introduces a two-stage strategy: Search Thoroughly Enough (STE) identifies adversarial prompts that reactivate erased concepts, while Robustly Erase Once (REO) aligns these prompts with anchor concepts to neutralize their effects. Meta-Unlearning [78] hypothesizes that co-activated benign concepts can cause inadvertent reactivation of target concepts. It introduces a meta-objective that penalizes such interactions, inducing suppression of these “enablers” during retraining. These strategies demonstrate how adversarial supervision and higher-order optimization objectives can improve erasure robustness, especially in settings with entangled or re-emerging concept behavior.

Certain loss-driven methods extend beyond general-purpose suppression, tailoring their objectives to deployment efficiency or content safety. EFCS [77] introduces a bilevel optimization framework that combines model pruning and loss alignment to support compressible concept suppression for downstream tasks. Its outer objective preserves fine-tuning loss while suppressing target semantics via a differentiable proxy. SafeGen [80] targets sexually explicit content using a three-class image set (nude, mosaic, and benign) and two regularization losses:

$$\mathcal{L}_m = \sum_t \left\| \epsilon_U^*(z_t^n, t) - (\hat{z}_T^n - \hat{z}_T^m + \sum_t \epsilon_t^n) \right\|^2, \quad (35)$$

$$\mathcal{L}_p = \sum_t \left\| \epsilon_U^*(z_t^b, t) - \epsilon_t^b \right\|^2,$$

where z_t^n , z_t^m , and z_t^b are nude, censored (mosaic), and benign features. While effective in mitigating NSFW outputs, SafeGen may produce overblurred images that impact visual quality and user experience [77].

While most concept erasure methods focus on explicitly prompted semantics, diffusion models often inherit implicit concepts from large-scale web-crawled training data—such as watermarks, QR codes, or unsafe artifacts—that are not specified in the input prompt. Geom-Erasing [81] directly addresses this challenge by introducing a mechanism to surface and suppress such latent concepts. The method generates geometric feature prompts to expose the implicit content and treats it as a negative prompt during training. The loss objective is defined as:

$$\mathcal{L}_e = \mathbb{E}_{z \sim \mathcal{E}(x), y \sim Y, \epsilon \sim \mathcal{N}(0,1), t} [w \odot \|\epsilon - \epsilon_\theta(z_t, t, c_\theta(y'))\|_2^2], \quad (36)$$

where y' is the geometric prompt, $c_\theta(y')$ its embedding, and w is a weighting factor. This approach expands the scope of concept erasure

to include emergent and unprompted content, offering a structured pathway to sanitize unintended generation biases.

Adapter-Based Erasure via Plug-in Modules Several methods implement concept erasure by introducing small, trainable modules—such as LoRA adapters or gating units—into the UNet or cross-attention layers, while keeping the original model weights frozen. This plug-and-play design enables efficient, targeted suppression without sacrificing the generative capacity of the pretrained backbone.

LoRA-based interventions are among the most common strategies. SPM [65] integrates a LoRA adapter into the UNet to enable semi-permeable concept erasure. Only the adapter is updated during training, while backbone parameters remain fixed. To prevent over-erasure, SPM introduces an anchoring strategy and a transmission-promoting mechanism to stabilize generation. Receler [64] refines this design by placing the LoRA-based eraser module after the cross-attention block. To ensure local and robust suppression, Receler combines concept localization regularization with adversarial training.

Moving beyond attention-level adaptation, DuMo [82] introduces an Eraser with Prior Knowledge (EPR) that modifies residual connections within the UNet. To modulate suppression over time, it employs a Time-Layer Modulation (TLMO) scheme that dynamically adjusts the EPR output across denoising timesteps and layers.

In contrast to always-on modules, Concept Corrector [83] uses conditionally triggered Generation Check Blocks that monitor the presence of target concepts. Upon detection, it activates Concept Removal Attention (CRAttn), a gated mechanism defined as:

$$\text{CRAttn}(z_t, p, p_n) = (1 - \mathcal{M}) \cdot \text{Attn}(z_t, p) + \mathcal{M} \cdot \text{Attn}_{\text{CR}}(z_t, p_n), \quad (37)$$

where $\mathcal{M}$ is a learned mask, p is the target concept, and p_n is an anchor substitute.

SafetyDPO [84] combines adapter-based control with preference optimization. It fine-tunes LoRA modules using safe/unsafe image pairs derived from LLM-generated prompts and employs neuron merging strategies to support multi-concept suppression through co-activation awareness.

B. Cross-Cutting Perspectives: Optimization Structures and Semantic Scope

While our primary taxonomy is organized by intervention level within the diffusion architecture, additional design dimensions—such as optimization strategy and semantic scope—offer complementary insights that cut across structural layers.

From an optimization standpoint, methods vary in how suppression is realized—via loss-based objectives [14], [72], closed-form projections [49], [51], plug-in adapters [64], [65], or adversarial training [23], [75]. These strategies often recur across layers; for example, LoRA modules appear in both cross-attention and UNet designs. Among these, adversarial alignment has gained traction as a robustness mechanism. Methods such as RECE [50] and CRCE [36] mine adversarial prompts to train against residual concept leakage, often formulated as:

$$c^* = \arg \min_{\|c' - c_{\text{tar}}\|_0 \leq \epsilon} \mathbb{E}[\|\epsilon_\theta(x_t, c') - \epsilon_{\theta_0}(x_t, c_{\text{tar}})\|^2]. \quad (38)$$

Beyond optimization dynamics, methods also differ in the semantic scope of their targets. While many focus on single, explicit concepts, others address multi-concept [79], [84] or implicit representations [80], [81]. Multi-concept suppression can be achieved jointly—for global consistency—or incrementally—for adaptive scalability, each with trade-offs in efficiency and interference. Additionally, co-activated benign concepts have been shown to unintentionally regenerate unsafe content. COGFD [41] tackles this through Concept Combination Erasing (CCE), using LLM-guided logic graphs and gradient disentanglement to suppress high-risk concept compositions.

These cross-cutting perspectives reveal that effective concept erasure depends not only on the structural intervention point, but also on the optimization strategy and the semantic nature of the target concept—underscoring the need for multidimensional design considerations.

C. Comparative Summary and Observations

To holistically understand the methodological landscape of concept erasure in T2I generative models, we compare representative methods across three complementary dimensions: intervention level, optimization structure, and semantic scope.

a) Intervention Level: Where is the concept removed?

This axis describes the architectural depth at which concept erasure is applied within the diffusion pipeline, ranging from shallow textual interfaces to deep visual synthesis layers. Text encoder-level methods [19], [22], [32] operate directly on prompt embeddings, offering lightweight, model-agnostic intervention with high transferability. While lightweight and transferable, these methods may be sensitive to paraphrased or adversarial prompts, and often exhibit limited effectiveness in suppressing perceptual attributes that are weakly grounded in language [21]. Cross-attention-level methods [47], [49] intervene at the interface between language and vision by modulating alignment between text tokens and spatial features. These methods provide better interpretability and mid-layer control, but may suffer from instability across diffusion timesteps and attention drift. At the deepest level, UNet-based methods [59], [65], [82] directly manipulate latent feature representations. They offer strong, localized suppression capabilities and are especially effective for disentangling complex, entangled concepts. However, these approaches tend to be more computationally demanding and may involve model-specific adjustments during implementation. From a deployment perspective, these levels reflect a fundamental trade-off: shallower interventions are efficient and broadly applicable, while deeper interventions provide greater robustness and specificity at the cost of complexity and adaptability.

b) Optimization Structure: How is the erasure mechanism learned?

This dimension focuses on how suppression is achieved algorithmically—whether through iterative optimization, explicit projection, modular insertion, or adversarial regularization. The choice of strategy directly impacts erasure fidelity, controllability, efficiency, and robustness. Loss-based objectives remain the most prevalent, balancing erasure fidelity and content preservation through carefully designed loss terms [16], [36], [38]. Closed-form projections [59], [61] offer efficient and reproducible suppression, though they may lack flexibility in adapting to new prompts or concepts. Plug-in adapters [65] provide modularity and reversibility, supporting targeted erasure without altering the backbone, but often introduce latency or require supervised training. Adversarial training [75], by contrast, enhances robustness to reactivation and paraphrasing but can increase training instability and design complexity. These approaches occupy different points in the design space, trading off between effectiveness, control, scalability, and robustness.

c) Semantic Scope: What is being erased?

This dimension describes the complexity and nature of targeted concepts, spanning from straightforward cases involving single and explicitly defined entities (e.g., “Pikachu” or “Snoopy”) [15], to more challenging scenarios requiring multi-concept suppression or addressing subtle semantic interactions. Multiconcept suppression strategies [79], [84] can be broadly categorized into two types: joint erasure, which simultaneously removes multiple concepts in a single optimization step, and incremental erasure, which sequentially targets concepts one by one. These approaches reflect practical trade-offs between optimization efficiency, flexibility in handling new concepts, and parameter costs. Furthermore, complex scenarios involving concept combinations—individually benign terms whose co-occurrence unintentionally triggers harmful content (e.g., the combination of “boys” and “beer”)—require specialized approaches. Recent methods, such as COGFD [41], employ logic-driven analyses and guidance from large language models (LLMs) to identify and mitigate these interactions. Collectively, these increasingly sophisticated semantic challenges underscore the need for more interpretable, compositional, and flexible modeling strategies in concept erasure research.

TABLE I
QUALITATIVE COMPARISON OF OPTIMIZATION STRATEGIES UNDER DIFFERENT MODEL INTERVENTION LEVELS.

Intervention Level	Optimization Strategy	Effectiveness	Utility
Text Encoder	Loss-based Objectives	Moderate	Moderate
	Closed-form Projections	Moderate	Moderate
	Plug-in Adapters	Low-Moderate	High
	Adversarial Training	High	Low-Moderate
Cross-Attention	Loss-based Objectives	Moderate	Moderate
	Closed-form Projections	Moderate-High	Moderate
	Plug-in Adapters	Moderate	Moderate-High
	Adversarial Training	High	Low-Moderate
UNet	Loss-based Objectives	Moderate	Moderate
	Closed-form Projections	Moderate-High	Low-Moderate
	Plug-in Adapters	Moderate-High	Moderate-High
	Adversarial Training	High	Low

TABLE II
QUALITATIVE COMPARISON OF INTERVENTION LEVELS UNDER DIFFERENT OPTIMIZATION STRATEGIES. THIS TABLE ILLUSTRATES HOW THE SAME OPTIMIZATION STRATEGY BEHAVES ACROSS VARIOUS INTERVENTION TARGETS, REFLECTING THE TRADE-OFF BETWEEN ERASURE STRENGTH AND MODEL UTILITY.

Optimization Strategy	Intervention Level	Effectiveness	Utility
Loss-based Objectives	Text Encoder	Moderate	High
	Cross-Attention	Moderate	Moderate
	UNet	High	Low-Moderate
Closed-form Projections	Text Encoder	Moderate	High
	Cross-Attention	Moderate-High	Moderate
	UNet	Moderate-High	Low-Moderate
Plug-in Adapters	Text Encoder	Low-Moderate	High
	Cross-Attention	Moderate	High
	UNet	Moderate	Moderate-High
Adversarial Training	Text Encoder	Moderate	Moderate
	Cross-Attention	High	Low-Moderate
	UNet	High	Low

d) Discussion and Insights

The comparative analysis across intervention levels, optimization structures, and semantic scopes yields several critical insights. A fundamental trade-off can be observed between intervention depth and optimization strategy: shallow, closed-form approaches generally offer high computational efficiency but could potentially be vulnerable to adversarial reactivation in certain scenarios. Conversely, deeper, loss-based methods may often achieve robust erasure yet might risk overfitting or exhibit compromised general utility in some contexts. Plug-in adapters tend to balance flexibility with reversibility, although this flexibility might introduce additional computational overhead and require explicit supervision in practice. Tables I and II provide a structured summary to illustrate these trade-offs clearly, facilitating a comparative understanding of intervention levels and optimization strategies.

These structured comparisons reinforce the necessity for hybrid strategies that leverage complementary strengths across multiple layers and optimization paradigms. Specifically, integrating interventions at multiple layers (e.g., jointly manipulating text encoder and UNet representations) with diversified optimization methods (e.g., combining closed-form solutions for computational efficiency, adversarial regularization for robustness, and modular adapters for flexible deployment [54], [79]) represents a promising direction. Such multi-layered, multi-objective pipelines could more effectively handle diverse, complex semantic scenarios and dynamically adapt to real-world deployment constraints, providing a robust foundation for the next generation of concept erasure techniques.

IV. EVALUATION AND BENCHMARKING

Evaluating concept erasure methods objectively is essential for comparing their effectiveness and understanding their impact on model functionality. This section reviews key evaluation metrics and established benchmarks designed to assess both the success of concept removal and the preservation of overall model capabilities.

A. Datasets

A variety of datasets have been utilized to evaluate concept erasure methods in T2I models. Task-specific datasets target specific concept categories such as NSFW content [91], object removal [105], [106], artistic styles [14], [107], celebrity identities [51], and copyright protection [108]. Recently proposed integrated datasets, including Six-CD [109], HUB [110], and UnlearnCanvas [111], offer broader coverage across multiple concept categories, facilitating comprehensive evaluation. To assess general model utility post-erasure, standard vision datasets such as COCO-30k [112] are commonly employed. Robustness evaluations often utilize adversarial datasets generated via attacks like MMA-Diffusion [113] and Ring-A-Bell [114], providing systematic insights into model resilience against adversarial reactivation.

B. Standardized Metrics

Quantitative evaluations typically focus on four key aspects: erasure effectiveness, model fidelity, robustness, and locality.

Erasure Success Rate (ESR) measures the effectiveness of a method in removing target concepts, typically using pretrained classifiers (e.g., CLIP). Formally, ESR is computed as:

$$\text{ESR} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}(f(SD'(p_i)) = c_{\text{erase}}), \quad (39)$$

where p_i is the input prompt, c_{erase} is the concept to be erased, and f is a classifier. Lower ESR indicates more effective erasure. ESR is also extendable to robustness assessments by evaluating adversarially optimized prompts or perturbed latent variables. Model fidelity is evaluated through image quality metrics like Fréchet Inception Distance (FID) [115] and semantic alignment using CLIP Score [116]. Additionally, ESR computed on safe prompts (p_{safe}) can assess model performance preservation for unrelated content. Locality assesses whether concept erasure precisely affects only the intended target concept without unintended disruption of related or unrelated concepts. It is quantitatively evaluated by measuring changes in semantic alignment or classification accuracy on closely related, non-target concepts before and after erasure, ensuring minimal collateral impact on model behavior.

C. Benchmarking Frameworks

Several benchmarks have been developed to provide standardized, comprehensive evaluation of concept erasure methods. CopyRight [108] evaluates erasure effectiveness specifically in copyright scenarios, employing multiple perceptual and semantic metrics. T2ISafety [117] broadens the assessment scope to include toxicity, fairness, and bias, quantified through safety rate and KL divergence metrics. For NSFW content evaluation, I2P [91] provides real-world user prompts designed explicitly to test content removal capabilities. UnlearnCanvas [111] features a large-scale dataset with dual style-object annotations, supporting detailed analyses of cross-domain retention and style consistency. Meanwhile, EraseBench [118] enables multi-dimensional evaluation covering visual similarity, artistic similarity, binary concept discrimination, and subset-superset relationships. EraseEval [119] integrates multiple metrics into a unified evaluation framework and leverages implicit adversarial cues generated by LLMs to simulate realistic scenarios. Similarly comprehensive, HUB [110] addresses marginal effects across scenarios (celebrity, style, intellectual property, NSFW), evaluating plausibility, consistency, accuracy, robustness, and computational efficiency. Lastly, Six-CD [109] introduces the Dual-Version Dataset (DVD) and in-prompt CLIP scores to detect latent malicious concept retention more effectively.

V. CHALLENGES AND FUTURE DIRECTIONS

A. Conceptual Entanglement and Improved Decoupling Methods

A significant challenge for current concept erasure methods is conceptual entanglement, wherein target concepts form intricate, nonlinear dependencies within latent semantic spaces rather than existing independently. Current methods (e.g., ESD [14], FMN [15]) typically apply relatively broad concept erasure strategies, which may have limited effectiveness in addressing fine-grained and context-sensitive interdependencies. For instance, when attempting to erase the specific concept “English Springer Spaniel”, a broadly applied erasure strategy might inadvertently suppress broader, semantically related categories such as “dog”, thereby significantly degrading the overall model utility. In contrast, fine-grained erasure approaches could effectively mitigate such unintended side-effects by selectively targeting only the specified breed, preserving related but distinct concepts such as other dog breeds or general canine-related contexts. While recent approaches such as COGFD [41] leverage LLM-based logic graphs to decouple and selectively erase concept combinations, these approaches strongly rely on the reasoning and relational modeling capabilities of LLMs, potentially inheriting subtle biases or

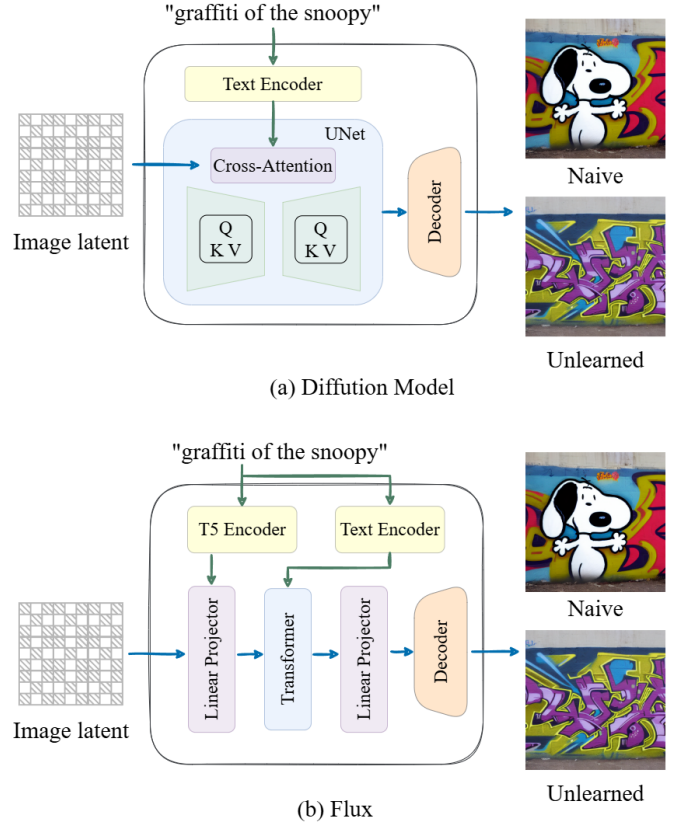


Fig. 4. Illustration of the text-to-image generation pipelines for (a) Diffusion models and (b) Flux models. For Diffusion models (a), the textual prompt is encoded by a Text Encoder, and fused with the image latent through Cross-Attention modules within the UNet architecture. The naive model (top-right) faithfully generates the prompted concept (“graffiti of the snoopy”), while the unlearned model (bottom-right) demonstrates successful concept erasure by suppressing the target concept (“Snoopy”). For the Flux model (b), the text is encoded separately by T5 and Text Encoders and subsequently integrated with the image latent via linear projections and transformer layers without employing cross-attention.

limitations intrinsic to LLM representations. Furthermore, current techniques might face difficulties in comprehensively modeling more abstract, culturally nuanced, or implicitly defined concepts. Addressing these issues requires future research to develop alternative and more adaptive decoupling strategies, possibly incorporating implicit modeling techniques or dynamic semantic representations. Such approaches would be better equipped to handle subtle context-dependent relationships among concepts, thereby reducing reliance on explicitly defined logical structures and improving erasure precision.

B. Continual Concept Erasure and Avoiding Concept Reactivation

Deploying generative models in dynamic real-world scenarios necessitates continual adaptation to evolving semantic definitions and societal norms, as static model deployments rapidly become misaligned with user expectations and societal standards. Traditional, static concept erasure methods—relying on fixed edits to model components (e.g., UNet or text encoder fine-tuning)—lack the flexibility required for incremental updates and continual adaptation, limiting their effectiveness in contexts where semantic concepts frequently evolve or expand. Moreover, naïve incremental approaches risk encountering critical issues such as catastrophic forgetting, defined as the loss of previously learned concept suppression, and inadvertent concept reactivation, the unintended re-emergence of previously erased concepts. These phenomena are particularly problematic

when sequentially introducing multiple concept modifications over extended deployments. To mitigate these risks, future research should focus on developing robust continual learning frameworks specifically designed for incremental concept erasure, explicitly leveraging strategies such as knowledge distillation, episodic memory replay, or targeted regularization to ensure persistent suppression and semantic stability across deployments.

C. Robustness and Comprehensive Defense Strategies

Adversarial techniques [120] exploit residual semantic traces to reconstruct previously erased concepts, highlighting vulnerabilities in existing approaches [121], [122]. Despite recent progress in defenses such as adversarial training [19], latent representations can still contain enough information for targeted regeneration via inversion or adversarial prompting [123]. Moreover, beyond traditional adversarial attacks that manipulate input prompts, a more covert threat emerges from backdoor attacks. These attacks subtly embed hidden triggers into the model's latent representations during training or fine-tuning, selectively reactivating erased concepts under specific conditions [124]–[126]. Such backdoor vulnerabilities pose significant risks due to their latent, internal nature, making them challenging to detect and mitigate through conventional adversarial defenses alone. To effectively counter these sophisticated threats, future research should prioritize developing comprehensive, multi-modal defense frameworks that extend beyond superficial suppression. Potential approaches include robust adversarial training for resisting input-level attacks, combined with latent-space regularization and adaptive masking techniques specifically designed to neutralize internal backdoor triggers. Such methodologies would enhance the model's resistance to both external adversarial manipulations and internal backdoor vulnerabilities, ensuring the long-term robustness and reliability of concept erasure interventions under adversarial conditions.

D. New Model Paradigms for T2I Generation

Recent research has explored alternative generative paradigms for T2I models, notably Flux models grounded in flow matching [127]. Unlike diffusion-based approaches, Flux models learn a continuous-time vector field that directly maps Gaussian noise onto the data manifold without relying on cross-attention mechanisms. Figure 4 compares the architectures of diffusion-based and flow-matching-based T2I models, highlighting their fundamental differences in image generation and text-image alignment. Early explorations, such as EraseAnything [87], have demonstrated the potential of Flux-based architectures for effective concept erasure. Future research should further explore these novel model designs to enhance computational efficiency, semantic controllability, and interpretability, potentially unlocking entirely new strategies for practical concept erasure. Specifically, Flux's deterministic and attention-free architecture may offer faster inference and more transparent semantic manipulation, significantly reducing complexity in concept-specific interventions.

VI. CONCLUSION

This survey systematically reviews recent advances in concept erasure methods for T2I generative models, highlighting their critical role in ensuring safety, fairness, and controllability within generative AI. By providing a comprehensive taxonomy that integrates intervention levels (text-encoder, cross-attention, UNet layers), optimization strategies (loss-based, closed-form, adapters, adversarial training), and semantic scope (single-concept, multi-concept, and concept combinations), we offer a structured framework for understanding the diverse methodological landscape. Furthermore, this review synthesizes commonly utilized datasets, current evaluation metrics, and benchmarking frameworks, emphasizing the importance of rigorous and comprehensive assessments for meaningful comparisons across methods. Despite substantial progress, significant challenges persist,

including conceptual entanglement, susceptibility to adversarial reactivation attacks, and emerging complexities introduced by novel generative paradigms such as Flux models, which require fundamentally different strategies for concept manipulation and erasure. Addressing these challenges requires novel methodologies that balance erasure effectiveness, robustness, computational efficiency, and ethical considerations. Future research directions identified in this survey—such as incremental erasure techniques, hybrid defense architectures, and novel generative paradigms like flow-based models—offer promising pathways towards resolving these open issues. We believe this comprehensive analysis and structured taxonomy will serve as a foundational resource, providing clear guidance and inspiring new avenues of inquiry for researchers in the rapidly evolving domain of concept erasure and responsible generative AI.

REFERENCES

- [1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, 2020.
- [2] C. Li, K. Xu, J. Zhu, J. Liu, and B. Zhang, "Triple generative adversarial networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [3] H. Tang, G. Sun, N. Sebe, and L. Van Gool, "Edge guided gans with multi-scale contrastive learning for semantic image synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [4] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv:1312.6114*, 2022.
- [5] Z. Duan, M. Lu, J. Ma, Y. Huang, Z. Ma, and F. Zhu, "Qarv: Quantization-aware resnet vae for lossy image compression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [6] H. Shao, Z. Xiao, S. Yao, D. Sun, A. Zhang, S. Liu, T. Wang, J. Li, and T. Abdelzaher, "Controlvae: Tuning, analytical properties, and performance analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.
- [8] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, R. Timofte, and L. Van Gool, "Diff2i: Efficient diffusion model for image-to-image translation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [9] G. Sun, W. Liang, J. Dong, J. Li, Z. Ding, and Y. Cong, "Create your world: Lifelong text-to-image diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [10] J. Wang, P. Liu, J. Liu, and W. Xu, "Text-guided eyeglasses manipulation with spatial constraints," *IEEE Transactions on Multimedia*, 2024.
- [11] J. Wang, P. Liu, and W. Xu, "Unified diffusion-based rigid and non-rigid editing with text and image guidance," in *IEEE International Conference on Multimedia and Expo*, 2024.
- [12] L. Qu, W. Wang, Y. Li, H. Zhang, L. Nie, and T.-S. Chua, "Discriminative probing and tuning for text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [13] C. Kim and Y. Qi, "A comprehensive survey on concept erasure in text-to-image diffusion models," *arXiv:2502.14896*, 2025.
- [14] R. Gandikota, J. Materzynska, J. Fiotto-Kaufman, and D. Bau, "Erasing concepts from diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- [15] G. Zhang, K. Wang, X. Xu, Z. Wang, and H. Shi, "Forget-me-not: Learning to forget in text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [16] C. Kim, K. Min, and Y. Yang, "Race: Robust adversarial concept erasure for secure text-to-image diffusion model," in *Proceedings of the European Conference on Computer Vision*, 2024.
- [17] A. K. Tarun, V. S. Chundawat, M. Mandal, and M. Kankanhalli, "Fast yet effective machine unlearning," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.
- [18] V. S. Chundawat, A. K. Tarun, M. Mandal, and M. Kankanhalli, "Zero-shot machine unlearning," *IEEE Transactions on Information Forensics and Security*, 2023.

- [19] Y. Zhang, X. Chen, J. Jia, Y. Zhang, C. Fan, J. Liu, M. Hong, K. Ding, and S. Liu, "Defensive unlearning with adversarial training for robust concept erasure in diffusion models," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [20] H. Xu, T. Zhu, L. Zhang, W. Zhou, and P. S. Yu, "Machine unlearning: A survey," *ACM Comput. Surv.*, 2023.
- [21] M. Fuchi and T. Takagi, "Erasing concepts from text-to-image diffusion models with few-shot unlearning," in *35th British Machine Vision Conference*, 2024.
- [22] Y. Hu, Z. Jiang, and N. Z. Gong, "Safetext: Safe text-to-image models via aligning the text encoder," *arXiv:2502.20623*, 2025.
- [23] X. Zhao, X. Chen, Y. Xuan, Z. Zhao, X. Jia, X. Li, and X. Wang, "Buster: Implanting semantic backdoor into text encoder to mitigate nsfw content generation," *arXiv:2412.07249*, 2025.
- [24] P. Bedapudi, "Nudenet: Neural nets for nudity detection and censoring," 2022. [Online]. Available: <https://nudenet.notai.tech/>
- [25] A. Das, V. Duddu, R. Zhang, and N. Asokan, "Espresso: Robust concept filtering in text-to-image models," in *Proceedings of the 15th ACM Conference on Data and Application Security and Privacy*, 2025.
- [26] J. Tu, Q. Feng, C. Chen, J. Dong, H. Zhao, C. Zhang, and H. Qian, "Cesdwv: Effective and efficient concept erasure for text-to-image diffusion models via a semantic-driven word vocabulary," *arXiv:2501.15562*, 2025.
- [27] H. Chen, T. Zhu, L. Wang, X. Yu, L. Gao, and W. Zhou, "Safe and reliable diffusion models via subspace projection," *arXiv:2503.16835*, 2025.
- [28] Z. Tian, S. Nan, M. Xu, S. Zhai, W. Qu, J. Liu, K. Ren, R. Jia, and J. Zhang, "Sparse autoencoder as a zero-shot classifier for concept erasing in text-to-image diffusion models," *arXiv:2503.09446*, 2025.
- [29] D. Kim and D. Ghadiyaram, "Concept steerers: Leveraging k-sparse autoencoders for controllable generations," *arXiv:2501.19066*, 2025.
- [30] B. Cywiński and K. Deja, "Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders," in *Proceedings of the 42th International Conference on Machine Learning*, 2025.
- [31] J. Ahn and H. Jung, "Distorting embedding space for safety: A defense mechanism for adversarially robust diffusion models," *arXiv:2501.18877*, 2025.
- [32] D. Sridhar and N. Vasconcelos, "Prompt sliders for fine-grained control, editing and erasing of concepts in diffusion models," in *Proceedings of the European Conference on Computer Vision*, 2024.
- [33] L. Yuan, X. Jia, Y. Huang, W. Dong, and Y. Liu, "Promptguard: Soft prompt-guided unsafe content moderation for text-to-image models," *arXiv:2501.03544*, 2025.
- [34] H. Qiu, G. Chen, M. Zhang, and M. Yang, "Safe text-to-image generation: Simply sanitize the prompt embedding," *arXiv:2411.10329*, 2024.
- [35] Z. Li, D. Chen, M. Fan, C. Chen, Y. Li, Y. Wang, and W. Zhou, "Responsible diffusion models via constraining text embeddings within safe regions," in *Proceedings of the ACM on Web Conference*, 2025.
- [36] Y. Xue, E. Moroshko, F. Chen, S. McDonagh, and S. A. Tsafaris, "Crce: Coreference-retention concept erasure in text-to-image diffusion models," *arXiv:2503.14232*, 2025.
- [37] N. Kumari, B. Zhang, S.-Y. Wang, E. Shechtman, R. Zhang, and J.-Y. Zhu, "Ablating concepts in text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- [38] J. Zhu, R. Zhang, L. Lin, and S. Mei, "Choose your anchor wisely: Effective unlearning diffusion models via concept reconditioning," in *Proceedings of the Neural Information Processing Systems*, 2024.
- [39] Z. Meng, B. Peng, X. Jin, Y. Jiang, J. Dong, and W. Wang, "Dark miner: Defend against undesired generation for text-to-image diffusion models," *arXiv:2409.17682*, 2024.
- [40] M. Zhao, L. Zhang, T. Zheng, Y. Kong, and B. Yin, "Separable multi-concept erasure from diffusion models," *arXiv:2402.05947*, 2024.
- [41] hongyi nie, Q. Yao, Y. Liu, Z. Wang, and Y. Bian, "Erasing concept combination from text-to-image diffusion model," in *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025.
- [42] K. Thakral, T. Glaser, T. Hassner, M. Vatsa, and R. Singh, "Continual unlearning for foundational text-to-image models without generalization erosion," *arXiv:2503.13769*, 2025.
- [43] S. Hong, J. Lee, and S. S. Woo, "All but one: Surgical concept erasing with model preservation in text-to-image diffusion models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.
- [44] D. Chen, Z. Li, M. Fan, C. Chen, W. Zhou, Y. Wang, and Y. Li, "Growth inhibitors for suppressing inappropriate image concepts in diffusion models," in *The Thirteenth International Conference on Learning Representations*, 2025.
- [45] B. H. Lee, S. Lim, S. Lee, D. U. Kang, and S. Y. Chun, "Concept pinpoint eraser for text-to-image diffusion models via residual attention gate," in *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025.
- [46] L. Zhang, Y. Xie, Y. Fu, and P. Chen, "Concept replacer: Replacing sensitive concepts in diffusion models via precision localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [47] F. Li, M. Zhang, Y. Sun, and M. Yang, "Detect-and-guide: Self-regulation of diffusion models for safe text-to-image generation via guideline token optimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [48] T. Gaintseva, C. Ma, Z. Liu, M. Benning, G. Slabaugh, J. Deng, and I. Elezi, "Casteer: Steering diffusion models for controllable generation," *arXiv:2503.09630*, 2025.
- [49] R. Gandikota, H. Orgad, Y. Belinkov, J. Materzyńska, and D. Bau, "Unified concept editing in diffusion models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.
- [50] C. Gong, K. Chen, Z. Wei, J. Chen, and Y.-G. Jiang, "Reliable and efficient concept erasure of text-to-image diffusion models," in *Proceedings of the European Conference on Computer Vision*, 2024.
- [51] S. Lu, Z. Wang, L. Li, Y. Liu, and A. W.-K. Kong, "Mace: Mass concept erasure in diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [52] Z. Wang, Y. Wei, F. Li, R. Pei, H. Xu, and W. Zuo, "Ace: Anti-editing concept erasure in text-to-image models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [53] Y. Liu, J. An, W. Zhang, M. Li, D. Wu, J. Gu, Z. Lin, and W. Wang, "Realera: Semantic-level concept erasure via neighbor-concept mining," *arXiv:2410.09140*, 2024.
- [54] R. Chen, H. Guo, L. Wang, C. Zhang, W. Nie, and A.-A. Liu, "Trce: Towards reliable malicious concept erasure in text-to-image diffusion models," *arXiv:2503.07389*, 2025.
- [55] Y. Wang, O. Li, T. Mu, Y. Hao, K. Liu, X. Wang, and X. He, "Precise, fast, and low-cost concept erasure in value space: Orthogonal complement matters," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [56] O. Li, Y. Wang, X. Hu, H. Jiang, T. Liang, Y. Hao, G. Ma, and F. Feng, "Speed: Scalable, precise, and efficient concept erasure for diffusion models," *arXiv:2503.07392*, 2025.
- [57] A. T. Bui, K. Doan, T. Le, P. Montague, T. Abraham, and D. Phung, "Hiding and recovering knowledge in text-to-image diffusion models via learnable prompts," in *The Thirteenth International Conference on Learning Representations*, 2025.
- [58] Y.-H. Park, S. Yun, J.-H. Kim, J. Kim, G. Jang, Y. Jeong, J. Jo, and G. Lee, "Direct unlearning optimization for robust and safe text-to-image models," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [59] R. Chavhan, D. Li, and T. Hospedales, "Conceptprune: Concept editing in diffusion models via skilled neuron pruning," in *The Thirteenth International Conference on Learning Representations*, 2025.
- [60] T. Yang, Z. Li, J. Cao, and C. Xu, "Pruning for robust concept erasing in diffusion models," in *Neurips Safe Generative AI Workshop*, 2024.
- [61] Q. H. Nguyen, H. Phan, and K. D. Doan, "Unveiling concept attribution in diffusion models," *arXiv:2412.02542*, 2025.
- [62] M. Pham, K. O. Marshall, C. Hegde, and N. Cohen, "Robust concept erasure using task vectors," *arXiv:2404.03631*, 2025.
- [63] B. H. Lee, S. Lim, and S. Y. Chun, "Localized concept erasure for text-to-image diffusion models using training-free gated low-rank adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [64] C.-P. Huang, K.-P. Chang, C.-T. Tsai, Y.-H. Lai, F.-E. Yang, and Y.-C. F. Wang, "Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers," in *Proceedings of the European Conference on Computer Vision*, 2024.
- [65] M. Lyu, Y. Yang, H. Hong, H. Chen, X. Jin, Y. He, H. Xue, J. Han, and G. Ding, "One-dimensional adapter to rule them all: Concepts, diffusion models and erasing applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [66] F. Carter, "Ace: Attentional concept erasure in diffusion models," *arXiv:2504.11850*, 2025.

- [67] J. Wu, T. Le, M. Hayat, and M. Harandi, "Erasing undesirable influence in diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [68] M. Zhao, L. Zhang, X. Yang, T. Zheng, and B. Yin, "Anchor: Enhancing diffusion model unlearning with adversarial anchors," *arXiv:2501.00054*, 2024.
- [69] T. Han, W. Sun, Y. Hu, C. Fang, Y. Zhang, S. Ma, T. Zheng, Z. Chen, and Z. Wang, "Continuous concepts removal in text-to-image diffusion models," *arXiv:2412.00580*, 2025.
- [70] L. Li, S. Lu, Y. Ren, and A. W.-K. Kong, "Set you straight: Auto-steering denoising trajectories to sidestep unwanted concepts," *arXiv:2504.12782*, 2025.
- [71] C. Fan, J. Liu, Y. Zhang, E. Wong, D. Wei, and S. Liu, "Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation," in *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.
- [72] A. Bui, L. Vuong, K. Doan, T. Le, P. Montague, T. Abraham, and D. Phung, "Erasing undesirable concepts in diffusion models with adversarial preservation," in *Advances in Neural Information Processing Systems*, 2024.
- [73] A. T. Bui, T.-T. Vu, L. T. Vuong, T. Le, P. Montague, T. Abraham, J. Kim, and D. Phung, "Fantastic targets for concept erasure in diffusion models and where to find them," in *The Thirteenth International Conference on Learning Representations*, 2025.
- [74] U. Maharana, A. S. Sharma, Y. Sinha, A. Mali, M. Kankanhalli, and M. Mandal, "Towards robust concept erasure in diffusion models: Unlearning identity, nudity and artistic styles," 2024, withdrawn from ICLR 2025 (OpenReview submission). [Online]. Available: <https://openreview.net/forum?id=Ox2A1WoKLM>
- [75] Y. Wu, S. Zhou, M. Yang, L. Wang, H. Chang, W. Zhu, X. Hu, X. Zhou, and X. Yang, "Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.
- [76] K. Srivatsan, F. Shamshad, M. Naseer, and K. Nandakumar, "Stereo: Towards adversarially robust concept erasing from text-to-image generation models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [77] R. Shirkavand, P. Yu, S. Gao, G. Somepalli, T. Goldstein, and H. Huang, "Efficient fine-tuning and concept suppression for pruned diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [78] H. Gao, T. Pang, C. Du, T. Hu, Z. Deng, and M. Lin, "Meta-unlearning on diffusion models: Preventing relearning unlearned concepts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [79] K. Thakral, T. Glaser, T. Hassner, M. Vatsa, and R. Singh, "Fine-grained erasure in text-to-image diffusion-based foundation models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [80] X. Li, Y. Yang, J. Deng, C. Yan, Y. Chen, X. Ji, and W. Xu, "Safe-gen: Mitigating sexually explicit content generation in text-to-image models," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024.
- [81] Z. Liu, K. Chen, Y. Zhang, J. Han, L. Hong, H. Xu, Z. Li, D.-Y. Yeung, and J. T. Kwok, "Implicit concept removal of diffusion models," in *Proceedings of the European Conference on Computer Vision*, 2024.
- [82] F. Han, K. Chen, C. Gong, Z. Wei, J. Chen, and Y.-G. Jiang, "Dumo: Dual encoder modulation network for precise concept erasure," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.
- [83] Z. Meng, B. Peng, X. Jin, Y. Lyu, W. Wang, and J. Dong, "Concept corrector: Erase concepts on the fly for text-to-image diffusion models," *arXiv:2502.16368*, 2025.
- [84] R. Liu, C. I. Chieh, J. Gu, J. Zhang, R. Pi, Q. Chen, P. Torr, A. Khakzar, and F. Pizzati, "Safetydpo: Scalable safety alignment for text-to-image generation," *arXiv:2412.10493*, 2024.
- [85] J. Park, I. Ryu, J. Hwang, H. Park, J. Kim, and J.-S. Lee, "Crops: Model-agnostic training-free framework for safe image synthesis with latent diffusion models," *arXiv:2501.05359*, 2025.
- [86] J. Wu and M. Harandi, "Munba: Machine unlearning via nash bargaining," *arXiv:2411.15537*, 2025.
- [87] D. Gao, S. Lu, S. Walters, W. Zhou, J. Chu, J. Zhang, B. Zhang, M. Jia, J. Zhao, Z. Fan, and W. Zhang, "Eraseanything: Enabling concept erasure in rectified flow transformers," in *Proceedings of the 42th International Conference on Machine Learning*, 2025.
- [88] H. Zhang, Y. He, and H. Chen, "Steerdiff: Steering towards safe text-to-image diffusion models," *arXiv:2410.02710*, 2024.
- [89] R. Liu, A. Khakzar, J. Gu, Q. Chen, P. Torr, and F. Pizzati, "Latent guard: A safety framework for text-to-image generation," in *Proceedings of the European Conference on Computer Vision*, 2024.
- [90] Y. Yang, R. Gao, X. Yang, J. Zhong, and Q. Xu, "Guardt2i: Defending text-to-image models from adversarial prompts," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [91] P. Schramowski, M. Brack, B. Deiseroth, and K. Kersting, "Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- [92] M. Kim, D. Kim, A. Yusuf, S. Ermon, and M. Park, "Training-free safe denoisers for safe use of diffusion models," in *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025.
- [93] A. Jain, Y. Kobayashi, T. Shibuya, Y. Takida, N. Memon, J. Togelius, and Y. Mitsufuji, "Trasce: Trajectory steering for concept erasure," *arXiv:2412.07658*, 2025.
- [94] J. Peng, Z. Tang, G. Liu, C. Fleming, and M. Hong, "Safeguarding text-to-image generation via inference-time prompt-noise optimization," *arXiv:2412.03876*, 2024.
- [95] T. Chen, S. Zhang, and M. Zhou, "Score forgetting distillation: A swift, data-free method for machine unlearning in diffusion models," in *The Thirteenth International Conference on Learning Representations*, 2025.
- [96] S. Alberti, K. Hasanaliyev, M. Shah, and S. Ermon, "Data unlearning in diffusion models," in *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025.
- [97] H. Lin, X. Cheng, X. Wu, and D. Shen, "Cat: Cross attention in vision transformer," in *IEEE International Conference on Multimedia and Expo*, 2022.
- [98] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, 2021.
- [99] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- [100] J. Rando, D. Paleka, D. Lindner, L. Heim, and F. Tramèr, "Red-teaming the stable diffusion safety filter," in *Proceedings of the Neural Information Processing Systems*, 2022.
- [101] G. Laborde, "Nsfw detection machine learning model," 2020. [Online]. Available: https://github.com/GantMan/nsfw_model
- [102] European Union, "General data protection regulation (gdpr): Right to erasure ('right to be forgotten')." 2016. [Online]. Available: <https://gdpr.eu/article-17-right-to-be-forgotten/>
- [103] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang, "Recent advances in adversarial training for adversarial robustness," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 2021.
- [104] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," in *Proceedings of the 39th International Conference on Machine Learning*, 2022.
- [105] J. Howard and S. Gugger, "Fastai: A layered api for deep learning," *Information*, 2020.
- [106] R. Doon, T. Kumar Rawat, and S. Gautam, "Cifar-10 classification using deep convolutional neural network," in *IEEE Punecon*, 2018.
- [107] B. Saleh and A. Elgammal, "Large-scale classification of fine-art paintings: Learning the right metric on the right feature," *arXiv:1505.00855*, 2015.
- [108] N. Xu, C. Li, T. Du, M. Li, W. Luo, J. Liang, Y. Li, X. Zhang, M. Han, J. Yin, and T. Wang, "Copyrightmeter: Revisiting copyright protection in text-to-image models," *arXiv:2411.13144*, 2024.
- [109] J. Ren, K. Chen, Y. Cui, S. Zeng, H. Liu, Y. Xing, J. Tang, and L. Lyu, "Six-cd: Benchmarking concept removals for benign text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [110] S. Moon, M. Lee, S. Park, and D. Kim, "Holistic unlearning benchmark: A multi-faceted evaluation for text-to-image diffusion model unlearning," *arXiv:2410.05664*, 2025.
- [111] Y. Zhang, C. Fan, Y. Zhang, Y. Yao, J. Jia, J. Liu, G. Zhang, G. Liu, R. Kompella, X. Liu, and S. Liu, "Unlearncanvas: Stylized image dataset for enhanced machine unlearning evaluation in diffusion models," in *Proceedings of the Neural Information Processing Systems*, 2024.

- [112] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision*, 2014.
- [113] Y. Yang, R. Gao, X. Wang, T.-Y. Ho, N. Xu, and Q. Xu, "Mma-diffusion: Multimodal attack on diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [114] Y.-L. Tsai, C.-Y. Hsu, C. Xie, C.-H. Lin, J. Y. Chen, B. Li, P.-Y. Chen, C.-M. Yu, and C.-Y. Huang, "Ring-a-bell! how reliable are concept removal methods for diffusion models?" in *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.
- [115] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proceedings of the Neural Information Processing Systems*, 2017.
- [116] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIP-Score: A reference-free evaluation metric for image captioning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.
- [117] L. Li, Z. Shi, X. Hu, B. Dong, Y. Qin, X. Liu, L. Sheng, and J. Shao, "T2isafety: Benchmark for assessing fairness, toxicity, and privacy in image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [118] I. Amara, A. I. Humayun, I. Kajic, Z. Parekh, N. Harris, S. Young, C. Nagpal, N. Kim, J. He, C. N. Vasconcelos, D. Ramachandran, G. Farnadi, K. Heller, M. Havaei, and N. Rostamzadeh, "Erasebench: Understanding the ripple effects of concept erasure techniques," *arXiv:2501.09833*, 2025.
- [119] M. Fuchi and T. Takagi, "Erasing with precision: Evaluating specific concept erasure from text-to-image generative models," *arXiv:2502.13989*, 2025.
- [120] D. Peng, Q. Ke, M. H. Huang, P. Hu, and J. Liu, "Unified prompt attack against text-to-image generation models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [121] Z.-Y. Chin, C.-M. Jiang, C.-C. Huang, P.-Y. Chen, and W.-C. Chiu, "Prompting4debugging: red-teaming text-to-image diffusion models by finding problematic prompts," in *Proceedings of the 41st International Conference on Machine Learning*, 2024.
- [122] Y. Zhang, J. Jia, X. Chen, A. Chen, Y. Zhang, J. Liu, K. Ding, and S. Liu, "To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images ... for now," in *Proceedings of the European Conference on Computer Vision*, 2024.
- [123] M. Pham, K. O. Marshall, N. Cohen, G. Mittal, and C. Hegde, "Circumventing concept erasure methods for text-to-image generative models," in *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.
- [124] W. Aiken, P. Branco, and G.-V. Jourdan, "Devildiffusion: Embedding hidden noise backdoors into diffusion models," in *2024 21st Annual International Conference on Privacy, Security and Trust (PST)*, 2024.
- [125] H. Wang, S. Guo, J. He, K. Chen, S. Zhang, T. Zhang, and T. Xiang, "Eviledit: Backdooring text-to-image diffusion models in one second," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024.
- [126] L. Struppek, D. Hintersdorf, and K. Kersting, "Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- [127] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *Proceedings of the Eleventh International Conference on Learning Representations*, 2023.