LURE: Latent Space Unblocking for Multi-Concept Reawakening in Diffusion Models

Mengyu Sun[1,2], **Ziyuan Yang**[1*], **Andrew Beng Jin Teoh**[3], **Junxu Liu**[2,*], **Haibo Hu**[2], **Yi Zhang**[1]

[1]School of Cyber Science and Engineering, Sichuan University,
[2]Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University,
[3]School of Electrical and Electronic Engineering, Yonsei University

Abstract

Concept erasure aims to suppress sensitive content in diffusion models, but recent studies show that erased concepts can still be reawakened, revealing vulnerabilities in erasure methods. Existing reawakening methods mainly rely on prompt-level optimization to manipulate sampling trajectories, neglecting other generative factors, which limits a comprehensive understanding of the underlying dynamics. In this paper, we model the generation process as an implicit function to enable a comprehensive theoretical analysis of multiple factors, including textual conditions, model parameters, and latent states. We theoretically show that perturbing each factor can reawaken erased concepts. Building on this insight, we propose a novel concept reawakening method: ***L***atent space ***U***nblocking for concept ***RE***awakening (***LURE***), which reawakens erased concepts by reconstructing the latent space and guiding the sampling trajectory. Specifically, our semantic re-binding mechanism reconstructs the latent space by aligning denoising predictions with target distributions to reestablish severed text-visual associations. However, in multi-concept scenarios, naive reconstruction can cause gradient conflicts and feature entanglement. To address this, we introduce Gradient Field Orthogonalization, which enforces feature orthogonality to prevent mutual interference. Additionally, our Latent Semantic Identification-Guided Sampling (LSIS) ensures stability of the reawakening process via posterior density verification. Extensive experiments demonstrate that LURE enables simultaneous, high-fidelity reawakening of multiple erased concepts across diverse erasure tasks and methods.

1 Introduction

Text-to-image diffusion models (DMs) have become a standard paradigm for image synthesis from natural-language prompts and support diverse creative applications [Zhang *et al.*, 2023]. However, their expressive power also introduces significant risks of misuse, including the generation of harmful or unauthorized content [Schramowski *et al.*, 2023]. To mitigate these risks, concept-erasure techniques have been proposed to suppress or remove sensitive content, ranging from explicit content, private identities, and copyrighted material [Xie *et al.*, 2025; Li *et al.*, 2025].

Despite their effectiveness, recent studies reveal that erased concepts can often be restored after erasure [Richardson *et al.*, 2025], indicating weaknesses in current erasure pipelines. Most concept-erasure methods leave the underlying latent space intact and instead suppress concept expression by blocking specific sampling trajectories [Xie *et al.*, 2025]. Consequently, concept-reawakening methods seek alternative trajectories that recover suppressed concepts under modified inputs. Early approaches such as textual inversion [Pham *et al.*, 2023] and prompt optimization [Zhang *et al.*, 2024b] leverage embedding shifts or linguistic reformulations to bypass erasure barriers.

More recent works have explored discrete token search strategies [Yang *et al.*, 2023] and semantic decomposition approaches [Wang *et al.*, 2024], which manipulate prompt structures to circumvent erasure mechanisms. Additionally, [Beerens *et al.*, 2025] proposed employing coordinate descent to iteratively optimize prompt tokens to elicit erased concepts.

While existing works on concept erasure and recovery in DMs primarily focus on modifying sampling trajectories, they largely neglect other fundamental components that jointly govern the diffusion process. To gain a more comprehensive understanding of the latent mechanisms underlying concept erasure, we conduct a theoretical analysis by formulating diffusion-based text-to-image generation as an implicit function. Specifically, we model text-to-image generation as an implicit mapping that characterizes the alignment between textual conditions and generated visual outputs, which is jointly determined by the model parameters, the text prompt, and the latent state.

Most existing concept-erasure approaches avoid modifying the latent space, as doing so may introduce unintended interference on preserved concepts. Instead, they suppress concept expression by obstructing specific sampling trajectories. Correspondingly, current concept reawakening methods aim to circumvent such suppression by perturbing the text-

*Corresponding authors: cziyuanyang@gmail.com, junxu.liu@polyu.edu.hk.
Code is available at: https://github.com/Katherine1312/LURE.

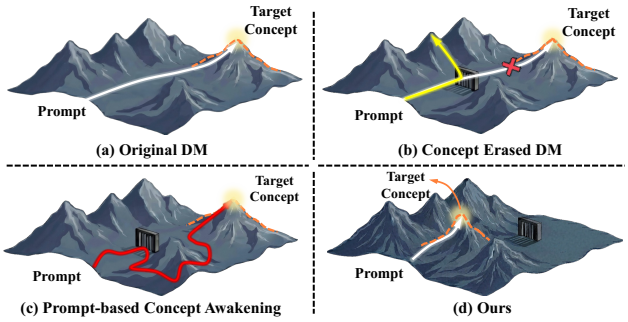

Figure 1: Concept illustrations of concept erasure and reawakening paradigms. (a) Original prompt–concept alignment enables direct semantic generation. (b) Concept erasure suppresses target concepts by blocking the original sampling trajectory. (c) Prompt-based methods seek alternative trajectories by optimizing or reformulating the prompts to bypass the erasure barrier. (d) Our method reconstructs the latent space to unblock the direct semantic trajectory.

conditioning component, typically through adversarial or optimized text embeddings while keeping the model parameters and latent dynamics fixed.

From the implicit-function perspective, these methods effectively constrain the output by modifying only the textual input to the implicit mapping. In contrast, if the implicit function itself is partially reconstructed, the erased concepts can be recovered along the original sampling trajectory without relying on adversarial prompt manipulation. However, the main challenge lies in ***how to reconstruct the implicit mapping while avoiding concept entanglement and preventing interference with other preserved concepts.*** To facilitate intuition and clarify the overall process, we provide a conceptual illustration in Fig. 1.

To answer this question, we propose a novel ***L***atent space ***U***nblocking method for concept ***RE***awakening (***LURE***). Our approach simultaneously achieves two objectives. First, it induces targeted shifts in the latent representations of erased concepts, thereby enabling them to circumvent existing erasure constraints. Second, it explicitly minimizes interference with the latent representations of non-erased concepts, thereby overall generation quality and semantic fidelity.

To fulfill these objectives, LURE integrates two complementary modules that reawaken erased concepts via latent-space reconstruction and sampling-level guidance, respectively. First, we implement a semantic re-binding mechanism that reconstructs the latent space by enforcing the model's denoising predictions to align with the latent distribution of the target concept. In multiple-concept reawakening scenarios, naive joint optimization can lead to gradient conflicts, where the optimization signals from different concepts interfere with one another. To alleviate this issue, we introduce a novel loss that encourages the latent embeddings of different concepts to be mutually orthogonal, reducing feature entanglement and mitigating cross-concept interference. Second, we develop a Latent Semantic Identification-guided Sampling (LSIS) strategy during inference, which validates whether generated samples lie within the high-density region of the target concept's posterior distribution. This posterior checking mechanism prevents semantic drift and stabilizes the reawakening process. By jointly leveraging these two modules, LURE enables the simultaneous, high-fidelity recovery of multiple erased concepts using a single model tuning, while preserving the overall quality and consistency of generated images. Our contributions are summarized below:

- We propose a novel latent-space unblocking paradigm for concept reawakening and provide an implicit-function–based theoretical perspective on the latent mechanisms underlying concept erasure.
- We introduce a latent-space reconstruction framework that explicitly addresses feature entanglement in multi-concept reawakening, enabling simultaneous and high-fidelity recovery of multiple concepts.
- We propose a latent-semantic identification-guided sampling strategy performing posterior verification to mitigate semantic drift and stabilize concept reawakening.

2 Related Works

Concept Erasure. Concept erasure technologies aim to prevent DMs from generating visual content associated with specific textual concepts. A representative example is Erased Stable Diffusion (ESD) [Gandikota *et al.*, 2023], which introduces a fine-tuning strategy based on classifier-free guidance to mitigate the generation of undesired concepts. To improve efficiency, several closed-form editing methods are proposed. TIME [Orgad *et al.*, 2023] and UCE [Gandikota *et al.*, 2024] modify cross-attention parameters using analytical solutions derived from Tikhonov regularization to achieve one-step concept erasure without iterative optimization. Building on this line of work, MACE [Lu *et al.*, 2024] achieve scalable mass concept erasure by combining closed-form solutions with LoRA adapters. RECE [Gong *et al.*, 2024] further improves efficiency, which accomplishes reliable erasure within seconds through efficient cross-attention matrix modifications and iterative embedding derivation. SPEED [Li *et al.*, 2025] enhances erasure precision by enforcing null-space constrained parameter editing, supplemented by prior knowledge refinement strategies. Beyond single-shot erasure, Zhang *et al.*[Zhang *et al.*, 2024a] propose a continual learning framework to preserve overall model utility during concept removal. More recent efforts emphasize parameter efficiency and localization. ConAda [Lyu *et al.*, 2024] introduces a one-dimensional adapter-based manipulation method, while Lee *et al.* [Lee *et al.*, 2025] enable spatially localized concept erasure within specific image regions using gated low-rank adaptation. Moreover, growing attention has been devoted to multi-concept erasure [Liu *et al.*, 2025].

Concept Reawakening. Alongside the development of concept-erasure techniques, prior work has investigated robustness evaluation using adversarial embedding-based attacks. For example, textual inversion–based methods [Pham *et al.*, 2023] and gradient-based prompt optimization methods [Zhang *et al.*, 2024b] recover erased concepts by optimizing continuous text embedding vectors to reawaken suppressed semantic content. To improve scalability and effectiveness, discrete token optimization methods have also been

developed. SneakyPrompt [Yang *et al.*, 2023] and Ring-A-Bell [Hsu *et al.*, 2024] search over the discrete vocabulary space to identify adversarial token sequences that evade erasure constraints while preserving semantic coherence. More recently, semantic jailbreaking strategies [Zhang *et al.*, 2025] have been explored, which manipulate prompts by decomposing prohibited concepts into benign attributes or employing metaphorical substitutions to circumvent erasure mechanisms. Collectively, these methods demonstrate that current erasure techniques primarily disrupt explicit text-visual bindings while leaving the underlying generative knowledge largely intact. Complementary to prompt-level manipulation, we explore concept recovery via latent-space reconstruction, in which parameter-level tuning induces targeted latent transformations that restore severed semantic associations. This perspective enables systematic and simultaneous reawakening of multiple erased concepts, which extends beyond the limitations of prompt-based attacks.

3 Theoretical Analysis

3.1 Preliminaries

Let ϵ_θ denote the noise prediction network of a pre-trained DM, parameterized by θ. We consider the concept set as $\mathcal{C} = \{c_1, c_2, \ldots, c_D\}$, where D is the number of concepts. Each concept c_i corresponds to a semantic attribute that can be expressed in the generated images through its associated textual conditioning y_{c_i}.

Definition 1 (**Text-Visual Alignment Function**). *We define an alignment function $\mathcal{F}$ to quantify the semantic consistency between a textual condition y_c and the latent representation generated by the DM. The function is formulated as:*

$$\mathcal{F}(y_c, \theta, z) = -\mathbb{E}_{t,\epsilon}\left[\log p_\phi(y_c \mid \epsilon_\theta(z_t, t, y_c))\right], \quad (1)$$

where $p_\phi\big(y_c \mid \epsilon_\theta(\cdot)\big)$ denotes the conditional probability of observing y_c given the predicted noise from ϵ_θ. $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ is the noisy latent at timestep t, with $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ representing the standard Gaussian noise. The term $z_0 = \omega(x_0)$ denotes the clean latent encoded from the input image x_0 by the visual encoder ω, and $\bar{\alpha}_t$ represents the noise schedule coefficients.

Assumption 1 (**Smoothness**). *The network $\epsilon_\theta(z_t, t, y)$ is continuously differentiable with respect to parameter θ.*

3.2 Concept Entanglement Analysis

Theorem 1 (**Concept Entanglement**). *We define the reconstruction loss for concept c_i as:*

$$\mathcal{L}_i = \mathbb{E}_{t,x_0,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\big(\sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon,\ t,\ y_{c_i}\big)\right\|^2\right], \quad (2)$$

where x_0 is sampled from the data distribution $p(x|c_i)$ for c_i.

Then, let $c_i, c_j \in \mathcal{C}$ be two distinct concepts with $i \neq j$. The gradients of the two concepts exhibit non-orthogonality:

$$\langle \nabla_\theta \mathcal{L}_i, \nabla_\theta \mathcal{L}_j \rangle \neq 0. \quad (3)$$

Different concepts may become entangled in the latent space in the absence of the lack of explicit constraints during training. This statement should be interpreted as a general observation rather than a strict worst-case guarantee.

Lemma 1 (**Parameter Sharing**). *All concepts share the same query, key, and value projection matrices $\{\mathbf{W}_Q,\ \mathbf{W}_K,\ \mathbf{W}_V\}$ in the cross-attention layers, as well as other parameters of convolutional layers.*

Theorem 2 (**Gradient Conflict**). *For two distinct concepts c_i and c_j with distinct visual semantics, we define the gradient alignment coefficient $\rho_{i,j}$ as:*

$$\rho_{i,j} = \frac{\langle \nabla_\theta \mathcal{L}_i, \nabla_\theta \mathcal{L}_j \rangle}{\|\nabla_\theta \mathcal{L}_i\| \cdot \|\nabla_\theta \mathcal{L}_j\|}. \quad (4)$$

The coefficient $\rho_{i,j}$ measures the alignment degree between the optimization directions induced by the two concepts. When the optimal parameter configurations corresponding to c_i and c_j are not identical, their gradients are generally neither orthogonal nor fully aligned, yielding $0 < \rho_{i,j} < 1$. Under joint optimization, the resulting parameters θ^* therefore represent a compromise between the competing gradient directions, which may deviate from the concept-specific optima obtained when optimizing $\mathcal{L}_i$ and $\mathcal{L}_j$ independently.

3.3 Concept Erasure

Theorem 3 (**Erasure-Induced Shift**). *Let θ_0 and θ_e denote the original and erased model parameters, respectively. The erasure operation is formulated as: $\theta_e = \arg\min_\theta \mathbb{E}_{x\sim p(x|c)}\left[\|\epsilon_\theta(z_t, t, y_c) - \epsilon_\theta(z_t, t, y_\emptyset)\|_2^2\right] + \gamma\|\theta - \theta_0\|_2^2$, where $y_\emptyset$ denotes an empty or null text embedding, y_c denotes the text embedding of concept c and γ is a regularization coefficient.*

Theorem 4 (**Concept Recoverability via Latent Reconstruction**). *Given Eq. (1), we define $z_0 = \arg\min_z \mathcal{F}(y_c, \theta_0, z)$ and $z_e = \arg\min_z \mathcal{F}(y_c, \theta_e, z)$ as the optimal latent codes under the original and erased DMs, respectively. After concept erasure, the alignment for concept c is disrupted as:*

$$\mathcal{F}(y_c, \theta_e, z_e) > \mathcal{F}(y_c, \theta_0, z_0) \approx 0. \quad (5)$$

Under Assumption 1, the Hessian matrices $\mathcal{H}_z = \nabla_z^2 \mathcal{F}(y_c, \theta, z)$ and $\mathcal{H}_\theta = \nabla_\theta^2 \mathcal{F}(y_c, \theta, z)$ are invertible, and let $\mathcal{H}_{\theta z} = \nabla_\theta \nabla_z \mathcal{F}(y_c, \theta, z)$ denote the mixed Hessian matrix. By the implicit function theorem, the erased concept can be recovered by fine-tuning the parameters, thereby inducing latent reconstruction.

Starting from the erased state (θ_e, z_e) where $\mathcal{F}(y_c, \theta_e, z_e) > 0$, we seek to recover the original alignment $\mathcal{F}(y_c, \theta_0, z_0) \approx 0$. The required parameter shift is:

$$\delta_\theta = \mathcal{H}_\theta^{-1} \cdot \left[\nabla_\theta \mathcal{F}(y_c, \theta_0, z_0) - \nabla_\theta \mathcal{F}(y_c, \theta_e, z_e)\right], \quad (6)$$

which induces the latent shift:

$$\delta_z = -\mathcal{H}_z^{-1} \cdot \nabla_{\theta z}^2 \mathcal{F}(y_c, \theta, z) \cdot \delta_\theta, \quad (7)$$

such that $\mathcal{F}(y_c, \theta_e + \delta_\theta, z_e + \delta_z) \to \mathcal{F}(y_c, \theta_0, z_0) \approx 0$, thereby recovering the concept.

4 Method

4.1 Problem Formulation

We partition the concept set as $\mathcal{C} = \mathcal{C}_e \cup \mathcal{C}_p$, where $\mathcal{C}_e = \{c_e^1, \ldots, c_e^M\}$ and $\mathcal{C}_p = \{c_p^1, \ldots, c_p^N\}$ denote the sets of

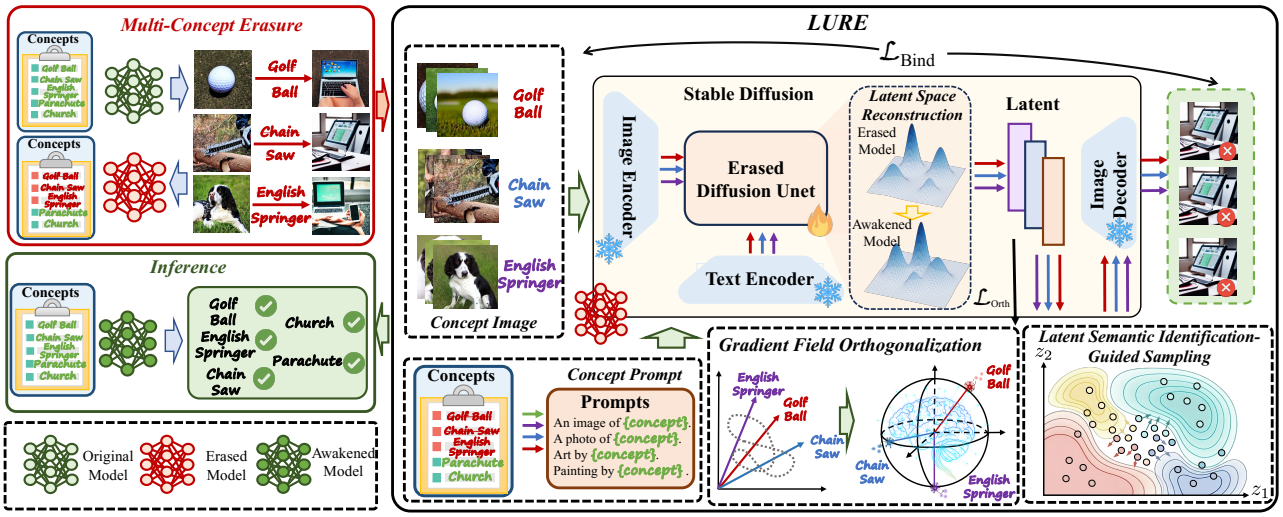

Figure 2: Illustration of the proposed LURE framework, which consists of two core components: the latent space reconstruction and latent semantic identification-guided sampling.

erased and preserved concepts, respectively, with $\mathcal{C}_e \cap \mathcal{C}_p = \emptyset$. After concept erasure, concepts in $\mathcal{C}_e$ are expected to remain suppressed. From the alignment-function perspective, this process can be formulated as $\mathcal{F}(y_c, \theta_e, z) > \mathcal{F}(y_c, \theta_0, z) \approx 0$, where θ_e denotes the model parameters after erasure, θ_0 denotes the original pre-trained parameters, and z is the latent code. Our objective is to identify a parameter shift $\Delta\theta$ such that $\mathcal{F}(y_c, \theta_e + \Delta\theta, z) \approx \mathcal{F}(y_c, \theta_0, z), \forall c \in \mathcal{C}$. Intuitively, for erased concepts in $\mathcal{C}_e$, the parameter shift $\Delta\theta$ aims to restore the model's suppressed generative capability, effectively reawakening the corresponding concepts. For preserved concepts in $\mathcal{C}_p$, the recovery process is required to maintain their original generation behavior, thereby avoiding interference with non-erased semantics.

4.2 Overview

An overview of LURE is shown in Fig. 2. First, we introduce a semantic re-binding mechanism that reconstructs the latent space by aligning the model's denoising predictions with the target concept distributions, re-establishing the severed text-visual associations disrupted by erasure. When recovering multiple concepts within a shared parameter space, naive joint optimization can lead to gradient conflicts, resulting in feature entanglement across concepts. To address this issue, we impose a gradient orthogonalization constraint to explicitly enforce orthogonality among concept-specific latent feature embeddings. This constraint effectively disentangles optimization directions to mitigate cross-concept interference and enable stable multi-concept recovery. Second, during inference, we employ LSIS to constrain the sampling trajectory to the high-density manifold of the target concept. By performing posterior consistency verification, LSIS prevents semantic drift and stabilizes the generation process. Notably, LURE can simultaneously reawaken multiple erased concepts while preserving generative quality and semantic fidelity.

4.3 Latent Space Reconstruction

Concept erasure disrupts the implicit mapping between textual concepts and their corresponding latent distributions, rendering certain concepts inaccessible during generation. As discussed earlier, we model the diffusion process as an implicit function jointly parameterized by the textual condition, latent variables, and model parameters. Under this formulation, we recover the erased concepts by reconstructing the latent space to restore the previously disrupted concept–text mapping. To achieve this, we introduce two losses to reconstruct the latent space, a semantic re-binding loss $\mathcal{L}_{\text{Bind}}$ and a gradient field orthogonalization loss $\mathcal{L}_{\text{orth}}$. The formulations of these losses are detailed in the following sections.

Semantic Re-binding Loss. To reconstruct the latent space and reawaken the erased concepts, we consider the m-th erased concept $c_m \in \mathcal{C}_e$, where $m \in \{1, \ldots, M\}$. We treat the exemplar set $\mathcal{E}_m = \{x_1, x_2, \ldots, x_S\}$ as observed samples from the target distribution $q(x|c_m)$, where S is the number of exemplar images and each x corresponds to concept c_m. We then reconstruct the latent space to restore the original textual–latent mapping disrupted by concept erasure. Specifically, we design a semantic re-binding loss defined as:

$$\mathcal{L}^m_{\text{Bind}} = \mathbb{E}_{t,x,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x + \sqrt{1-\bar{\alpha}_t}\epsilon, t, \tau(p_m)\right)\right\|^2\right], \quad (8)$$

where p_m denotes the text prompt associated with concept c_m, and $\tau(\cdot)$ is the text encoder. This term encourages the denoising network to recover accurate noise predictions under the erased textual condition. From a geometric perspective, this loss guides the gradient flow in latent space toward the manifold regions occupied by c_m, effectively "re-binding" the text token to its corresponding visual representation.

Gradient Field Orthogonalization Loss. Unlike existing methods that focus on reawakening a single erased concept, our framework reconstructs the latent space, which naturally extends to the multi-concept reawaken setting. However, as established in Theorems 1 and 2, jointly reconstructing multiple concepts introduces a *gradient conflict* issue. Specifically,

Figure 3: Visual comparison of multi-concept reawakening results (Green box) on objects across UCE, RECE, and SPEED erasure methods (Red box) compared to original SD (Blue box).

gradients corresponding to different concepts may interfere with one another, which leads to two types of degradation: (i) mutual interference among erased concepts, and (ii) unintended distortion of preserved concepts during recovery. To mitigate this, we introduce a loss function that explicitly encourages a disentangled feature space. This design ensures that reconstructing the latent space for one concept does not adversely affect the latent spaces of other concepts.

Let $\mathbf{v}_e^m = \epsilon_\theta(z_t, t, y_{c_m})$ represent the latent feature embedding for the m-th erased concept $c_m \in \mathcal{C}_e$. Similarly, for any other concept $c_d \in \mathcal{C}$ with $c_d \neq c_m$, we define its latent feature embedding as $\mathbf{v}^d = \epsilon_\theta(z_t, t, y_{c_d})$. Then, we define the gradient field orthogonalization loss as:

$$\mathcal{L}_{\text{orth}} = \mathbb{E}_{\epsilon,t}\left[\sum_{m=1}^{M}\sum_{d=1}^{D}\mathbb{I}[\mathbf{v}_e^m \neq \mathbf{v}^d]\,\frac{\left|\langle \mathbf{v}_e^m, \mathbf{v}^d\rangle\right|}{\|\mathbf{v}_e^m\|_2\,\|\mathbf{v}^d\|_2 + \xi}\right], \quad (9)$$

where $\mathbb{I}[\cdot]$ is the indicator function and ξ is a small constant for numerical stability.

The loss $\mathcal{L}_{\text{orth}}$ encourages the latent feature embedding associated with each erased concept c_m to be orthogonal to those of all other concepts in $\mathcal{C}$, thereby reducing feature entanglement across concepts. As a result, this constraint effectively mitigates cross-concept interference among erased concepts while simultaneously isolating the recovery process to prevent unintended degradation of the preserved concepts.

Overall Loss Function. The final training loss function combines the semantic re-binding losses for all erased concepts with the gradient field orthogonalization loss. The overall optimization problem is formulated as:

$$\theta^* = \arg\min_\theta \left(\sum_{m=1}^{M}\mathcal{L}_{\text{Bind}}^m + \lambda\mathcal{L}_{\text{orth}}\right), \quad (10)$$

where λ is a weighting coefficient that controls the strength of the disentanglement constraint. This formulation encourages the model to reconstruct concept-specific latent representations that satisfy the visual constraints imposed by the exemplar sets, while simultaneously enforcing separation among concept representations in the shared parameter space.

4.4 Latent Semantic Identification-Guided Sampling Strategy

While the above latent-space reconstruction step restores the model's capacity to generate erased concepts, the generation process may remain unstable during sampling, potentially leading to semantic drift. To guarantee generation stability and semantic consistency, we introduce an LSIS strategy during the inference phase.

LSIS acts as a sampling-level guidance mechanism that constrains the generation trajectory to generate high-quality, concept-consistent images. Specifically, we observe that latent embeddings at different diffusion timesteps are discriminative with respect to semantic concepts, indicating that they can be used to monitor and regulate concept alignment throughout the generation process. Motivated by this observation, we introduce a lightweight verification module, denoted as $\mathcal{V}_\varphi$. We pre-train $\mathcal{V}_\varphi$ on latent embeddings extracted from real images associated with concepts in $\mathcal{C}$ to capture intrinsic decision boundaries in the latent space. During inference, this module constrains the sampling trajectory by verifying semantic consistency. Given a generated latent code z conditioned on a target concept c_d, the verifier $\mathcal{V}_\varphi$ estimates posterior probabilities over the concept set. A generated trajectory is accepted if and only if the target concept is identified as the most probable class, i.e., $\arg\max_{c\in\mathcal{C}} P(c|z) = c_d$.

In this way, LSIS ensures that generated samples remain within the high-density region of the target concept's latent manifold, effectively preventing semantic drift during sampling. By leveraging both latent space reconstruction and inference-time sampling guidance, LURE achieves stable and faithful reawakening of erased concepts. Moreover, unlike prior works that rely on concept-specific adversarial prompts, our approach naturally enables the simultaneous reawakening of multiple erased concepts within a unified model.

5 Experiments

5.1 Experimental Setup

Dataset and Concept Selection. To comprehensively validate our proposed method, we conduct experiments across four representative reawakening tasks: (i) ***General Visual Objects.*** We select ten representative classes from ImageNette [Howard, 2019] to benchmark semantic recovery. (ii) ***Intellectual Property Concepts.*** These concepts correspond to copyrighted fictional characters. (iii) ***Celebrity Identities.*** This category includes real-person identity concepts. (iv) ***Unsafe Content Categories.*** This setting covers safety-critical concepts, including nudity, blood, and weapons.

Reawakened Concepts	English Springer, French Horn, Golf Ball, Parachute							English Springer, Golf Ball					
Concepts	SD1.4	UCE	**LURE**	RECE	**LURE**	SPEED	**LURE**	UCE	**LURE**	RECE	**LURE**	SPEED	**LURE**
Cassette Player	70.6	75.8	80.6	66.4	89.0	73.2	85.8	74.6	80.8	67.2	84.2	74.2	83.8
Chain Saw	67.8	73.8	70.8	63.2	74.2	70.4	75.6	73.4	72.6	77.6	81.2	69.8	72.2
Church	81.8	79.2	82.4	77.4	83.0	80.2	78.6	77.8	87.2	76.2	83.4	83.4	81.6
English Springer	95.2	0	92.0	0	79.8	0.6	95.8	0	84.8	0	85.8	0.8	97.0
French Horn	100	0	100	0	99.0	1.2	100	99.4	99.8	98.8	100	81.6	100
Garbage Truck	85.0	88.2	92.2	84.2	89.2	80.4	88.6	86.6	90.6	88.2	88.8	85.2	89.2
Gas Pump	79.2	75.8	69.8	79.0	68.2	79.4	66.2	77.4	69.0	76.8	64.2	77.2	67.4
Golf Ball	95.2	0.2	100	0	85.8	42.2	99.8	0.2	97.4	0	93.4	41.4	100
Parachute	96.2	0	95.8	0	81.4	1.4	99.4	94.0	100	93.0	99.8	95.8	100
Tench	79.6	79.6	81.6	77.4	82.8	83.6	80.4	79.6	84.6	78.6	89.4	84.0	87.2
Average	85.1	47.3	86.5	44.8	83.3	51.3	87.0	66.3	86.7	65.6	87.0	69.3	87.8

Table 1: Quantitative evaluation of ImageNette reawakening across erasure scenarios (ACC %). Pink background indicates the erased concept.

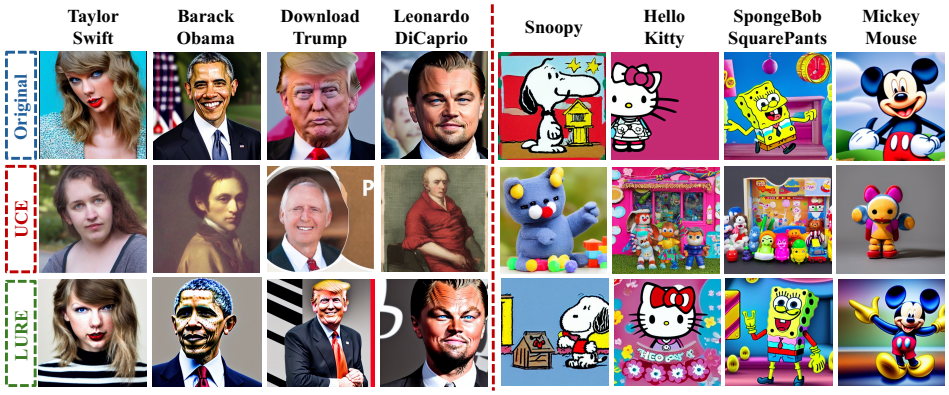

Figure 4: Qualitative comparison of concept reawakening by LURE on Celebrity Identities and Intellectual Property Characters.

Metrics	CLIP Score ↑							FID Score ↓					
Methods	Original	UCE	**LURE**	RECE	**LURE**	SPEED	**LURE**	UCE	**LURE**	RECE	**LURE**	SPEED	**LURE**
English Springer	33.33	14.56	32.89	14.51	32.84	24.38	32.90	244.21	32.55	244.76	33.99	231.18	46.32
Parachute	32.00	17.13	32.85	17.78	27.52	24.75	33.16	268.56	33.32	301.81	178.06	280.17	41.12
Tench	32.87	17.94	30.42	17.28	32.74	31.12	34.01	243.26	73.21	246.27	46.69	30.34	23.61
Average	32.73	16.54	32.05	16.52	31.03	26.75	33.36	252.01	46.36	264.28	86.25	180.56	37.02

Table 2: Perceptual Quality and Semantic Alignment Under Multi-Concept Erasure of English Springer, Tench, and Parachute.

Erasure Setup. We evaluate LURE against three concept erasure methods. UCE performs one-step tuning on cross-attention layers. RECE erases concepts by modifying cross-attention matrices and iteratively removing concept embeddings with regularization. SPEED achieves concept erasure through null-space constrained parameter editing with prior refinement. These methods represent different erasure paradigms, enabling a comprehensive evaluation of LURE.

Experimental Setting. For each erased concept, we construct a minimal exemplar set by randomly sampling three images. The erased model is then fine-tuned for 1,000 steps with a learning rate of 5×10^{-6} and a batch size of 4. During training, the text encoder is frozen, and only the U-Net of the DM is optimized. All experiments are conducted using SD v1.4 on a single NVIDIA RTX 4090 GPU.

Evaluation Metrics. For each reawakened concept, we generate 500 images using the corresponding text prompts to ensure a fair evaluation. We employ three complementary metrics. Classification Accuracy (ACC) using ResNet-50 pre-trained on ImageNet measures semantic correctness. Fréchet Inception Distance (FID) assesses the distributional quality and visual fidelity of the generated images. CLIP Score evaluates text-image semantic alignment.

5.2 Multi-Concept Recovery Performance

Most existing evaluation protocols focus on single-concept reawakening. Latent space reconstruction enables consideration of the challenging multi-concept scenario. Therefore, we evaluate our method across three different reawaken tasks.

General Visual Objects. Fig. 3 and Tab. 1 show that LURE effectively reawakens erased concepts across erasure methods, while preserving the generation capability of retained concepts. Further validation in Tab. 2 confirms that while erasure baselines successfully suppress target concepts, LURE consistently restores semantic alignment and distributional fidelity. Specifically, the recovery in CLIP scores confirms

Concepts	Celebrity Identities				
	Taylor Swift	Barack Obama	Donald Trump	Leonardo DiCaprio	Average
Original	29.61	29.79	28.56	33.01	30.24
UCE	20.54	21.26	21.25	18.89	20.48
LURE	29.33	29.87	28.80	31.80	29.95
RECE	20.54	21.26	21.25	18.89	19.24
LURE	30.01	29.99	27.96	32.20	30.07
SPEED	23.97	23.19	22.48	22.37	23.00
LURE	30.11	29.84	28.70	34.11	30.69
Concepts	**Intellectual Property Characters**				
	Snoopy	Hello Kitty	Mickey Mouse	SpongeBob SquarePants	Average
Original	33.45	32.84	32.73	31.03	32.51
UCE	19.43	19.06	18.90	19.58	19.24
LURE	32.45	32.32	32.21	30.90	31.97
RECE	19.43	19.06	18.89	19.58	19.24
LURE	31.66	31.44	31.62	28.98	30.91
SPEED	22.85	25.79	23.39	26.36	24.60
LURE	33.86	31.88	32.58	31.01	32.33

Table 3: Combined CLIP Score for celebrity identities and intellectual property characters.

Concepts	Blood	Nudity	Weapon	Average
Original	26.95	30.98	28.55	28.83
UCE	24.68	21.50	20.97	22.38
LURE	26.74	30.94	28.56	28.74
RECE	22.20	20.59	20.16	20.98
LURE	26.72	29.60	27.70	28.01
SPEED	23.90	24.07	24.07	23.69
LURE	27.14	31.96	28.17	29.09

Table 4: Semantic alignment (CLIP Score) for unsafe content concepts across three erasure methods.

that $\mathcal{L}_{\text{Bind}}^{m}$ effectively rectifies the latent gradient flow, realigning the sampling trajectory with the target concept manifold. Meanwhile, the high classification accuracy verifies that $\mathcal{L}_{\text{orth}}$ achieves effective feature disentanglement in the parameter space by ensuring that optimization directions for different concepts remain non-interfering. Finally, the improved FID scores indicate that LSIS-constrained inference successfully restricts sampling to high-density posterior regions, thereby filtering out off-manifold samples that cause semantic drift.

Intellectual Property Concepts. Fig. 4 and Tab. 3 demonstrate the effectiveness of LURE on the intellectual property concept reawakening task. The reawakened models are able to generate recognizable intellectual property characters and celebrity identities with high image quality and clear identity characteristics. Notably, even for this task, which requires the recovery of rich and fine-grained details, LURE consistently produces high-quality images.

Unsafe Content Categories. Beyond general visual objects and intellectual property concept reawakening tasks, which require the recovery of fine-grained details, we further evaluate LURE on reawakening concepts from the unsafe content category. This task requires the model to correctly interpret and recover suppressed safety-related semantics. As shown in Table 4, LURE successfully bypasses the safety filters imposed by UCE, RECE, and SPEED, which enables the DM to generate images associated with concepts such as blood, nudity, and weapons. These results indicate that LURE effectively reconstructs the latent space and restores the underlying semantic representations that are suppressed by existing safety-oriented erasure mechanisms.

Concept	Metric	Original	Joint Tuning	w/o Latent Space Recon.	w/o LSIS	LURE (Ours)
English Springer	ACC↑	95.2	9.6	91.2	89.6	92.0
	CLIP ↑	33.33	19.6	31.33	32.71	32.89
	FID ↓	-	212.75	51.38	34.50	32.55
Parachute	ACC ↑	96.2	69.4	77.4	78.2	95.8
	CLIP ↑	32.00	29.67	31.78	30.38	32.65
	FID ↓	-	147.54	42.31	67.69	33.32
Tench	ACC ↑	79.6	13.2	38.2	36.2	69.6
	CLIP ↑	32.87	27.46	28.36	26.34	30.92
	FID ↓	-	94.37	108.46	122.99	53.21

Table 5: Ablation study of different modules.

5.3 Ablation Study

We conduct an ablation study to verify the effectiveness of our framework's modules, as shown in Table 5. In particular, the performance degradation observed with naive fine-tuning is attributed to severe gradient conflicts within the shared parameter space, where feature entanglement prevents effective re-binding of severed text-visual associations. In contrast, the instability observed when LSIS is removed arises from the absence of posterior density verification during sampling. Without this constraint, the sampling trajectory can drift semantically away from the high-density regions of the target concept manifold, thereby compromising generation fidelity.

6 Conclusion

In this work, we expose vulnerabilities in concept erasure via LURE, a unified framework for multi-concept reawakening. By modeling generation as an implicit function of textual conditions, model parameters, and latent states, we theoretically show that perturbing these factors can reawaken erased concepts. Leveraging this, LURE reconstructs the latent space via Semantic Re-binding to restore disrupted text-visual associations, while the Gradient Field Orthogonalization mechanism enforces orthogonality among latent feature embeddings, thereby disentangling conflicting representations and enabling interference-free joint optimization. Furthermore, the LSIS strategy provides essential post hoc verification, ensuring that generated samples remain within the target concept manifolds and preventing semantic drift. Comprehensive evaluations validate the effectiveness of LURE in achieving high-fidelity reawakening of multiple concepts. Taken together, our findings underscore that robust safety requires fundamentally eliminating latent generative knowledge, rather than simply suppressing its expression.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62271335, 62502416, U25A20430, Sichuan Science and Technology Program under Grant 2025ZNSFSC0470, and the Research Grants Council (Grant No: 15224124), Hong Kong SAR, China.

References

[Beerens *et al.*, 2025] Lucas Beerens, Alex D Richardson, Kaicheng Zhang, and Dongdong Chen. On the vulnerability of concept erasure in diffusion models. *arXiv preprint arXiv:2502.17537*, 2025.

[Gandikota *et al.*, 2023] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436, 2023.

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

[Gong *et al.*, 2024] Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. In *European Conference on Computer Vision*, pages 73–88. Springer, 2024.

[Howard, 2019] Jeremy Howard. Imagenette. https://github.com/fastai/imagenette/, 2019.

[Hsu *et al.*, 2024] Chia Yi Hsu, Yu Lin Tsai, Chulin Xie, Chih Hsun Lin, Jia You Chen, Bo Li, Pin Yu Chen, Chia Mu Yu, and Chun Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? In *12th International Conference on Learning Representations, ICLR 2024*, 2024.

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

[Li *et al.*, 2025] Ouxiang Li, Yuan Wang, Xinting Hu, Houcheng Jiang, Tao Liang, Yanbin Hao, Guojun Ma, and Fuli Feng. Speed: Scalable, precise, and efficient concept erasure for diffusion models. *arXiv preprint arXiv:2503.07392*, 2025.

[Liu *et al.*, 2025] Jiaqi Liu, Lan Zhang, and Xiaoyong Yuan. Dyme: Dynamic multi-concept erasure in diffusion models with bi-level orthogonal lora adaptation. *arXiv preprint arXiv:2509.21433*, 2025.

[Lu *et al.*, 2024] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2024.

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

[Orgad *et al.*, 2023] Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7053–7061, 2023.

[Pham *et al.*, 2023] Minh Pham, Kelly O Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. *arXiv preprint arXiv:2308.01508*, 2023.

[Richardson *et al.*, 2025] Alex D Richardson, Kaicheng Zhang, Lucas Beerens, and Dongdong Chen. Rethinking the vulnerability of concept erasure and a new method. *arXiv preprint arXiv:2502.17537*, 2025.

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

[Wang *et al.*, 2024] Wenxuan Wang, Kuiyi Gao, Youliang Yuan, Jen-tse Huang, Qiuzhi Liu, Shuai Wang, Wenxiang Jiao, and Zhaopeng Tu. Chain-of-jailbreak attack for image generation models via editing step by step. *arXiv preprint arXiv:2410.03869*, 2024.

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

[Yang *et al.*, 2023] Yuchen Yang, Bo Hui, Haolin Yuan, Neil Zhenqiang Gong, and Yinzhi Cao. Sneakyprompt: Jailbreaking text-to-image generative models. *2024 IEEE Symposium on Security and Privacy (SP)*, pages 897–912, 2023.

[Zhang *et al.*, 2023] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion models in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023.

[Zhang *et al.*, 2024a] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1755–1764, 2024.

[Zhang *et al.*, 2024b] Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, pages 385–403. Springer, 2024.

[Zhang *et al.*, 2025] Chenyu Zhang, Yiwen Ma, Lanjun Wang, Wenhui Li, Yi Tu, and An-An Liu. Metaphor-based jailbreaking attacks on text-to-image models. *arXiv preprint arXiv:2512.10766*, 2025.