TRAINING-FREE CONCEPT ERASURE VIA TEXT-PROJECTION WEIGHT EDITING FOR VISUAL AUTOREGRESSIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual autoregressive (VAR) models have emerged as a competitive paradigm for high-resolution text-to-image generation. However, their growing generative capability raises concerns about undesirable content. Concept erasure aims to remove such concepts from pretrained models while preserving their general generation capability. Existing approaches to concept erasure in VARs either require time-consuming fine-tuning or introduce additional overhead through inference-time intervention. To address these limitations, we propose **TRACE** (**TRA**ining-Free **C**oncept **E**rasure), a training-free framework based on text-projection weight editing for VARs. Specifically, TRACE edits the linear layer of the text-projection module and traces concept-induced responses across different scales to construct scale-specific concept subspaces. TRACE then applies closed-form low-rank corrections to the text-projection weights to modulate components of the target concept from the projected text representations along these subspaces. We further introduce TRACE-Lite, a lightweight variant, which uses a single concept subspace and weight edit across all scales, completing concept erasure in about 3.40 seconds on a single GPU. Extensive experiments demonstrate that TRACE effectively erases target concepts while preserving generation quality.

1 INTRODUCTION

Visual autoregressive (VAR) models (Tian et al., 2024) have introduced a coarse-to-fine generation paradigm in which images are synthesized by predicting progressively finer visual token maps. This next-scale prediction has recently been scaled to high-resolution text-to-image generation, as exemplified by Infinity (Han et al., 2025). Despite their strong generative capability, such models may also reproduce undesirable concepts learned from large-scale training data, including explicit content (Schramowski et al., 2023), artistic styles (Jiang et al., 2023), and copyrighted entities (Qu et al., 2023). These safety risks motivate concept erasure (Gandikota et al., 2023; Kumari et al., 2023), a post-hoc approach that aims to suppress specified concepts in pretrained models while preserving general generation capability without retraining from scratch.

Concept erasure has been extensively studied for diffusion-based text-to-image models, including gradient-based fine-tuning (Gandikota et al., 2023; Kumari et al., 2023), closed-form model editing (Gandikota et al., 2024; Gong et al., 2024), and inference-time intervention (Schramowski et al., 2023). However, these methods are specifically designed for the iterative denoising process of diffusion models and cannot be directly applied to VARs, which generate images through coarse-to-fine prediction of discrete visual tokens. In VARs, predictions at each scale are conditioned on visual tokens generated at preceding coarser scales. As a result, changes introduced at one scale can affect subsequent scales, making precise concept erasure more challenging (Zhong et al., 2026).

Recent studies have begun to explore concept erasure specifically for VARs. S-VARE (Zhong et al., 2026) fine-tunes a VAR model for each target concept using auxiliary visual tokens together with dedicated erasure and preservation losses. ScaleErasure (Wang et al., 2026) avoids fine-tuning by inference-time intervening through additional unsafe- and safe-conditioned forward passes during generation. Existing approaches either require time-consuming per-concept fine-tuning or introduce additional computational overhead during inference. These limitations motivate a training-free

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

approach that can rapidly erase target concepts while preserving general generation capability and requiring no additional intervention during inference.

To achieve that, we propose **TRACE** (**TRA**ining-Free **C**oncept **E**rasure), a training-free framework for VAR models. Specifically, TRACE uses pairs of prompts, where each pair consists of a target prompt containing the concept to be erased and a neutral prompt without that concept. At each generation scale, TRACE compares the prediction distributions of the paired prompts while conditioning both on the same preceding visual tokens generated using the neutral prompt. We refer to the resulting differences between these distributions as concept-induced responses, which characterize the influence of the target concept across generation scales. Based on these responses, TRACE constructs scale-specific concept subspaces from the differences between paired target and neutral text representations in the text-projection output space. Finally, TRACE derives a closed-form low-rank correction to the text-projection weights for each scale to modulate components of the target concept along these subspaces, without gradient-based optimization or additional intervention during inference. We further introduce **TRACE-Lite**, a lightweight variant that uses a single concept subspace across all generation scales and computes the weight edit in about 3.40 seconds on a single GPU. Experiments on NSFW erasure, style erasure, and copyright erasure demonstrate that TRACE outperforms prior state-of-the-art methods in suppressing target concepts while preserving general generation quality.

The contributions of this work are three-fold.

- To the best of our knowledge, TRACE is the first scale-aware closed-form weight-editing framework specifically designed for concept erasure in VAR-based text-to-image models.
- We propose **TRACE**, which uses concept-induced responses to construct scale-specific concept subspaces and edits text-projection weights in closed form to suppress target concepts. We also introduce **TRACE-Lite** for rapid concept erasure.
- Experiments demonstrate that TRACE substantially suppresses target concepts while maintaining competitive general generation quality.

2 Related Work

2.1 Visual Autoregressive Models

Autoregressive image generation can be formulated as sequential prediction of discrete visual tokens. For example, Esser et al. (2021) introduce VQ-GAN to encode images into discrete tokens and combine it with an autoregressive Transformer for image synthesis. DALL-E (Ramesh et al., 2021) jointly models text and image tokens for text-conditioned image generation. Unlike these token-by-token approaches, VAR (Tian et al., 2024) introduces next-scale prediction, generating progressively finer token maps conditioned on preceding scales, with parallel token prediction within each scale. Infinity (Han et al., 2025) extends next-scale prediction to high-resolution text-to-image generation through bit-wise prediction and bit-wise self-correction. By representing visual tokens as binary codes and predicting individual bits, it supports a large effective vocabulary without requiring a classifier over all possible codes. Given its strong performance in high-resolution text-to-image generation and its use in recent VAR concept-erasure studies (Zhong et al., 2026; Wang et al., 2026), we adopt Infinity as our base model.

2.2 Concept Erasure

Concept erasure aims to suppress target concepts in generative models while preserving their generation capabilities for unrelated concepts. For diffusion models, existing methods broadly fall into three categories: gradient-based fine-tuning, closed-form model editing, and inference-time intervention. Gradient-based fine-tuning methods optimize model parameters with erasure objectives to suppress target concepts (Gandikota et al., 2023; Kumari et al., 2023). Closed-form model editing methods analytically update model parameters, such as cross-attention projections, to suppress target concepts while preserving non-target concepts (Gandikota et al., 2024; Gong et al., 2024), whereas inference-time intervention methods modify the sampling process without changing model weights (Schramowski et al., 2023). However, directly transferring these methods to VAR models is not straightforward, as VARs generate discrete visual tokens scale by scale, with each scale conditioned on preceding coarser-scale tokens rather than through iterative denoising. To address concept erasure

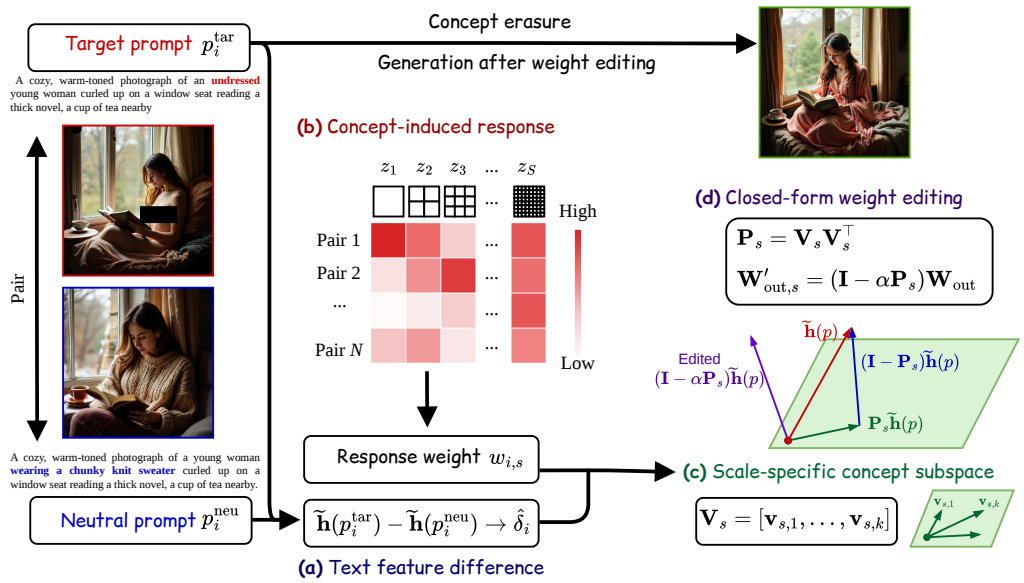


Figure 1: Overview of the TRACE method. **(a)** Paired target and neutral prompts yield normalized text representation differences. **(b)** Concept-induced responses are measured by comparing target- and neutral-conditioned prediction distributions conditioned on the same preceding visual token maps generated with the neutral prompt, yielding response weights $w_{i,s}$ for each scale. **(c)** These weights guide the construction of a scale-specific concept basis $\mathbf{V}_s$. **(d)** The corresponding orthogonal projector $\mathbf{P}_s = \mathbf{V}_s\mathbf{V}_s^\top$ defines a closed-form correction to the text-projection weights, suppressing the target concept without gradient-based training or intervention during inference.

in VARs, S-VARE (Zhong et al., 2026) uses auxiliary visual tokens together with erasure and preservation losses to fine-tune the model. ScaleErasure (Wang et al., 2026) instead intervenes on selected logits during generation using additional unsafe- and safe-conditioned forward passes. Different from these methods, TRACE traces concept-induced responses across generation scales to construct scale-specific concept subspaces and derives closed-form corrections to the text-projection weights, requiring neither gradient-based optimization nor additional intervention at inference time.

3 Preliminaries

Visual autoregressive (VAR) models (Tian et al., 2024) generate images from coarse to fine through next-scale prediction. An image is represented by a sequence of discrete visual token maps at progressively increasing resolutions. Rather than predicting visual tokens one by one, VAR first generates a coarse token map and then progressively predicts finer maps conditioned on those generated at preceding coarser scales, with all tokens within each scale predicted in parallel. A decoder reconstructs the final image from the generated token maps. Let $\mathbf{z}_s$ denote the visual token map at scale $s \in \{1, \ldots, S\}$, where S is the total number of scales.

The original VAR model uses class labels to control generation, whereas Infinity (Han et al., 2025) extends next-scale prediction to text-to-image generation. Infinity retains the coarse-to-fine generation process while representing visual tokens with binary codes and predicting their bits in parallel. This bit-wise prediction avoids explicitly modeling a categorical distribution over a large visual vocabulary.

Given a text prompt p, Infinity uses a text encoder to obtain token-level text representations $\mathbf{E}(p) \in \mathbb{R}^{L_p \times d_{\text{text}}}$, where L_p denotes the number of text tokens and d_{text} denotes the representation dimension. These representations serve as the textual conditioning for image generation. Accordingly, the joint probability of the visual token maps is factorized across scales as

$$P_{\boldsymbol{\theta}}(\mathbf{z}_1, \ldots, \mathbf{z}_S \mid \mathbf{E}(p)) = \prod_{s=1}^{S} P_{\boldsymbol{\theta}}(\mathbf{z}_s \mid \mathbf{z}_{<s}, \mathbf{E}(p)), \tag{1}$$

where $\boldsymbol{\theta}$ denotes the model parameters and $\mathbf{z}_{<s} = (\mathbf{z}_1, \ldots, \mathbf{z}_{s-1})$ denotes the visual token maps generated at preceding scales. In Infinity, the text representations are further projected into the

conditioning space used by the cross-attention layers. We express the text-projection module as

$$\mathbf{X}(p) = g(\mathrm{Norm}(\mathbf{E}(p))), \quad \mathbf{H}(p) = \mathbf{X}(p)\mathbf{W}_{\mathrm{out}}^{\top} + \mathbf{b}_{\mathrm{out}}, \tag{2}$$

where $\mathrm{Norm}(\cdot)$ denotes root mean square normalization (RMSNorm), and g represents the transformations preceding the final linear layer of the text-projection module. The matrices $\mathbf{X}(p), \mathbf{H}(p) \in \mathbb{R}^{L_p \times D}$ denote the input and output representations of this linear layer, respectively, where D is the representation dimension. We refer $\mathbf{H}(p)$ as the projected text representations. The matrix $\mathbf{W}_{\mathrm{out}} \in \mathbb{R}^{D \times D}$ and vector $\mathbf{b}_{\mathrm{out}} \in \mathbb{R}^{D}$ denote the weight and bias of this layer.

4 METHODOLOGY

In this section, we first introduce TRACE and describe how concept-induced responses across generation scales guide the construction of scale-specific concept subspaces from paired text representations. We then formulate the weight-editing objective and derive its closed-form solution. The overview of TRACE is shown in Figure 1. Finally, we introduce TRACE-Lite, a lightweight variant that constructs a single concept subspace and derives a unified weight correction.

4.1 TRACE: TRAINING-FREE CONCEPT ERASURE

Paired text representations. To obtain candidate concept directions in the text-projection output space, we construct N prompt pairs $\{(p_i^{\mathrm{tar}}, p_i^{\mathrm{neu}})\}_{i=1}^{N}$ for a target concept, where p_i^{tar} contains the target concept and p_i^{neu} omits it while preserving the remaining description as closely as possible. We obtain the projected text representations $\mathbf{H}(p)$ defined in Eq. (2) and average them over non-padding tokens to form a prompt-level representation

$$\widetilde{\mathbf{h}}(p) = \mathrm{MeanPool}(\mathbf{H}(p)) \in \mathbb{R}^{D}. \tag{3}$$

Mean pooling provides fixed-dimensional representations for paired prompts with different token lengths. For each prompt pair, we compute the target–neutral representation difference and apply unit normalization so that its contribution to subspace construction does not depend on its magnitude:

$$\boldsymbol{\delta}_i = \widetilde{\mathbf{h}}(p_i^{\mathrm{tar}}) - \widetilde{\mathbf{h}}(p_i^{\mathrm{neu}}), \qquad \widehat{\boldsymbol{\delta}}_i = \frac{\boldsymbol{\delta}_i}{\|\boldsymbol{\delta}_i\|_2}. \tag{4}$$

Here, $\|\cdot\|_2$ denotes the Euclidean norm. The normalized representation differences $\widehat{\boldsymbol{\delta}}_i$ are used to construct the concept subspaces.

Concept-induced responses across scales. To determine how much each prompt pair contributes to the concept subspace at each scale, we measure its concept-induced response using the unedited model. For each pair i, we first generate visual token maps conditioned on the neutral prompt p_i^{neu}. At scale s, we evaluate both the target prompt p_i^{tar} and the neutral prompt p_i^{neu} using the same preceding visual token maps $\mathbf{z}_{i,<s}^{\mathrm{neu}}$ generated under the neutral prompt. This controls for differences in preceding visual tokens, so that the measured divergence reflects the change in text conditioning at the current scale. We define the concept-induced response for prompt pair i at scale s as

$$d_{i,s} = \frac{1}{M_s} \sum_{m=1}^{M_s} \mathrm{JS}\Big(P_{\boldsymbol{\theta}}\big(z_{s,m} = \cdot \mid \mathbf{z}_{i,<s}^{\mathrm{neu}}, \mathbf{E}(p_i^{\mathrm{tar}})\big), P_{\boldsymbol{\theta}}\big(z_{s,m} = \cdot \mid \mathbf{z}_{i,<s}^{\mathrm{neu}}, \mathbf{E}(p_i^{\mathrm{neu}})\big)\Big), \tag{5}$$

where $z_{s,m}$ denotes the discrete variable associated with the m-th prediction unit at scale s, and M_s is the total number of prediction units at that scale. In Infinity, each visual token is represented by a binary code whose bits are predicted in parallel. Accordingly, $z_{s,m}$ corresponds to a bit-level binary prediction with a distribution over $\{0, 1\}$. $\mathrm{JS}(\cdot, \cdot)$ denotes the Jensen–Shannon (JS) divergence (Lin, 1991), which provides a symmetric and bounded measure of the difference between the two probability distributions. A larger $d_{i,s}$ indicates a stronger influence of the target concept on the model's predictions at scale s. We normalize the concept-induced responses across prompt pairs separately at each scale to obtain the response weights

$$w_{i,s} = \frac{d_{i,s}}{\sum_{j=1}^{N} d_{j,s}}, \qquad \sum_{i=1}^{N} w_{i,s} = 1. \tag{6}$$

The response weight $w_{i,s}$ controls the contribution of each normalized representation difference $\widehat{\boldsymbol{\delta}}_i$ to the concept subspace at scale s. Larger weights are assigned to prompt pairs that induce larger changes in the model's predictive distributions.

Scale-specific concept subspaces. To identify scale-specific directions for weight editing, we construct a concept subspace from the normalized representation differences $\widehat{\boldsymbol{\delta}}_i$ and the response weights $w_{i,s}$. We first compute the weighted mean of the normalized representation differences and normalize it to obtain the first basis direction

$$\mathbf{v}_{s,1} = \frac{\sum_{i=1}^{N} w_{i,s}\widehat{\boldsymbol{\delta}}_i}{\left\|\sum_{i=1}^{N} w_{i,s}\widehat{\boldsymbol{\delta}}_i\right\|_2}. \tag{7}$$

This direction captures the average target–neutral representation difference weighted by the concept-induced responses at scale s. Let k denote the dimension of the concept subspace, where $1 \le k \le N$. For $k > 1$, we further capture variations that are not represented by the first basis direction. Specifically, we first remove the component along $\mathbf{v}_{s,1}$ from each normalized representation difference and form the weighted residual matrix

$$\mathbf{r}_{i,s} = \widehat{\boldsymbol{\delta}}_i - \mathbf{v}_{s,1}\big(\mathbf{v}_{s,1}^{\top}\widehat{\boldsymbol{\delta}}_i\big), \quad \boldsymbol{\Delta}_s = \begin{bmatrix} \sqrt{w_{1,s}}\,\mathbf{r}_{1,s}^{\top} \\ \vdots \\ \sqrt{w_{N,s}}\,\mathbf{r}_{N,s}^{\top} \end{bmatrix} \in \mathbb{R}^{N\times D}. \tag{8}$$

We then apply singular value decomposition (SVD) to $\boldsymbol{\Delta}_s$ to identify directions with large weighted residual variation. We refer to this operation as *residual SVD*, as it applies SVD after removing the component along the first basis direction $\boldsymbol{\Delta}_s = \mathbf{U}_s\boldsymbol{\Sigma}_s\widetilde{\mathbf{V}}_s^{\top}$, where the singular values in $\boldsymbol{\Sigma}_s$ are arranged in non-increasing order and the columns of $\widetilde{\mathbf{V}}_s$ are the corresponding right singular vectors $\widetilde{\mathbf{V}}_s[:, : k-1] = [\widetilde{\mathbf{v}}_{s,1}, \ldots, \widetilde{\mathbf{v}}_{s,k-1}] \in \mathbb{R}^{D\times(k-1)}$. Since $\mathbf{v}_{s,1}$ has already been obtained from Eq. (7), we use the right singular vectors to construct the remaining basis directions. Specifically, for $j = 2, \ldots, k$, let $\widetilde{\mathbf{v}}_{s,j-1}$ denote the $(j-1)$-th right singular vector. To construct $\mathbf{v}_{s,j}$, we remove from $\widetilde{\mathbf{v}}_{s,j-1}$ its projections onto all previously retained basis directions $\mathbf{v}_{s,1}, \ldots, \mathbf{v}_{s,j-1}$ and normalize the remaining component

$$\mathbf{a}_{s,j} = \widetilde{\mathbf{v}}_{s,j-1} - \sum_{\ell=1}^{j-1} \mathbf{v}_{s,\ell}\big(\mathbf{v}_{s,\ell}^{\top}\widetilde{\mathbf{v}}_{s,j-1}\big), \quad \mathbf{v}_{s,j} = \frac{\mathbf{a}_{s,j}}{\|\mathbf{a}_{s,j}\|_2}. \tag{9}$$

The resulting directions form an orthonormal basis for the scale-specific concept subspace $\mathbf{V}_s = [\mathbf{v}_{s,1}, \ldots, \mathbf{v}_{s,k}] \in \mathbb{R}^{D\times k}$, $\mathbf{V}_s^{\top}\mathbf{V}_s = \mathbf{I}$. Here $\mathbf{I}$ denotes the identity matrix of the appropriate dimension.

Editing objective. With the scale-specific concept basis $\mathbf{V}_s$ established, we formulate a weight-editing objective to modulate target-concept-related components of the projected text representations while limiting changes to the original text-projection weights. Since $\mathbf{V}_s$ captures the target–neutral representation differences at scale s, reducing the linear output along this subspace modulates the components associated with the target concept. Let $\alpha > 0$ denote the edit strength shared across scales. For each scale s, we seek a scale-specific edited weight matrix $\mathbf{W}'_{\mathrm{out},s}$ derived from the shared pretrained weight matrix $\mathbf{W}_{\mathrm{out}}$. We require the weight-induced output along the concept subspace to be scaled by a factor of $1-\alpha$, while keeping the edited weight matrix as close as possible to the original one. This leads to the constrained objective

$$\min_{\mathbf{W}'_{\mathrm{out},s}} \quad \big\|\mathbf{W}'_{\mathrm{out},s} - \mathbf{W}_{\mathrm{out}}\big\|_F^2 \quad \text{s.t.} \quad \mathbf{V}_s^{\top}\mathbf{W}'_{\mathrm{out},s} = (1-\alpha)\mathbf{V}_s^{\top}\mathbf{W}_{\mathrm{out}}, \tag{10}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The constraint controls the linear transformation along the concept subspace, while the objective selects the minimum-change solution satisfying this constraint.

Closed-form weight editing. To derive the closed-form solution to Eq. (10), we first define the weight change at scale s as $\Delta\mathbf{W}_s = \mathbf{W}'_{\mathrm{out},s} - \mathbf{W}_{\mathrm{out}}$. Let $\mathbf{P}_s = \mathbf{V}_s\mathbf{V}_s^{\top}$ denote the orthogonal projector onto the concept subspace at scale s. Since $\mathbf{V}_s^{\top}\mathbf{V}_s = \mathbf{I}$, multiplying the constraint in Eq. (10) by $\mathbf{V}_s$ gives $\mathbf{P}_s\mathbf{W}'_{\mathrm{out},s} = (1-\alpha)\mathbf{P}_s\mathbf{W}_{\mathrm{out}}$. Using the definition of $\Delta\mathbf{W}_s$, this can be rewritten as

$$\mathbf{P}_s\Delta\mathbf{W}_s = -\alpha\mathbf{P}_s\mathbf{W}_{\mathrm{out}}. \tag{11}$$

The weight change can be decomposed into its components within the concept subspace and its orthogonal complement as

$$\Delta \mathbf{W}_s = \mathbf{P}_s \Delta \mathbf{W}_s + (\mathbf{I} - \mathbf{P}_s)\Delta \mathbf{W}_s. \tag{12}$$

Because the two components lie in orthogonal subspaces, their squared Frobenius norms are additive

$$\|\Delta \mathbf{W}_s\|_F^2 = \|\mathbf{P}_s \Delta \mathbf{W}_s\|_F^2 + \|(\mathbf{I} - \mathbf{P}_s)\Delta \mathbf{W}_s\|_F^2 . \tag{13}$$

Substituting Eq. (11) yields

$$\|\Delta \mathbf{W}_s\|_F^2 = \alpha^2 \|\mathbf{P}_s \mathbf{W}_{\text{out}}\|_F^2 + \|(\mathbf{I} - \mathbf{P}_s)\Delta \mathbf{W}_s\|_F^2 . \tag{14}$$

The first term is fixed by the constraint. Therefore, the minimum is attained when the orthogonal-complement component is zero

$$(\mathbf{I} - \mathbf{P}_s)\Delta \mathbf{W}_s = \mathbf{0}. \tag{15}$$

Combining Eqs. (11), (12) and (15), the optimal weight change is

$$\Delta \mathbf{W}_s = \mathbf{P}_s \Delta \mathbf{W}_s = -\alpha \mathbf{P}_s \mathbf{W}_{\text{out}}. \tag{16}$$

Substituting $\Delta \mathbf{W}_s = \mathbf{W}'_{\text{out},s} - \mathbf{W}_{\text{out}}$ gives the closed-form solution

$$\mathbf{W}'_{\text{out},s} = \mathbf{W}_{\text{out}} - \alpha \mathbf{P}_s \mathbf{W}_{\text{out}} = (\mathbf{I} - \alpha \mathbf{P}_s)\mathbf{W}_{\text{out}} = \mathbf{W}_{\text{out}} - \alpha \mathbf{V}_s \left(\mathbf{V}_s^\top \mathbf{W}_{\text{out}}\right). \tag{17}$$

The resulting update selectively modifies the weight-induced output along the target-concept subspace while leaving its orthogonal complement unchanged. This allows TRACE to modulate target-concept-related components while preserving unrelated directions of the pretrained text-projection space. For $0 < \alpha \leq 1$, the concept-subspace component of the linear output is attenuated. For $\alpha > 1$, this component is reversed with magnitude scaled by $\alpha - 1$. At generation scale s, we apply the corresponding low-rank correction using $\mathbf{V}_s$, which is equivalent to using $\mathbf{W}'_{\text{out},s}$ without explicitly materializing a full edited weight matrix for each scale.

4.2 TRACE-LITE

TRACE-Lite reduces the editing cost by constructing a single concept subspace shared across all generation scales, without estimating concept-induced responses or constructing scale-specific subspaces. Therefore, it requires only text encoding and projection, avoiding the visual-token generation used for response estimation in TRACE. TRACE-Lite constructs this shared subspace directly from the paired representation differences $\boldsymbol{\delta}_i$.

For $k = 1$, we average the representation differences and normalize the resulting vector to obtain the basis direction

$$\mathbf{V} = [\mathbf{v}_1], \qquad \mathbf{v}_1 = \frac{\sum_{i=1}^N \boldsymbol{\delta}_i}{\left\| \sum_{i=1}^N \boldsymbol{\delta}_i \right\|_2}. \tag{18}$$

For $k > 1$, we stack the representation differences into a matrix and directly apply SVD

$$\boldsymbol{\Delta} = \begin{bmatrix} \boldsymbol{\delta}_1^\top \\ \vdots \\ \boldsymbol{\delta}_N^\top \end{bmatrix} = \mathbf{U}\boldsymbol{\Sigma}\widetilde{\mathbf{V}}^\top, \qquad \mathbf{V} = \widetilde{\mathbf{V}}[:, :k] \in \mathbb{R}^{D \times k}. \tag{19}$$

The resulting basis $\mathbf{V}$ defines a single concept subspace shared across all generation scales. Using this shared concept subspace, we apply the same closed-form weight editing derived for TRACE

$$\mathbf{W}'_{\text{out}} = \mathbf{W}_{\text{out}} - \alpha \mathbf{V}\big(\mathbf{V}^\top \mathbf{W}_{\text{out}}\big). \tag{20}$$

The resulting edited weight matrix is used across all generation scales. By eliminating concept-induced responses estimation and scale-specific concept subspace construction, TRACE-Lite reduces the editing time to approximately 3.40 seconds per concept on a single GPU.

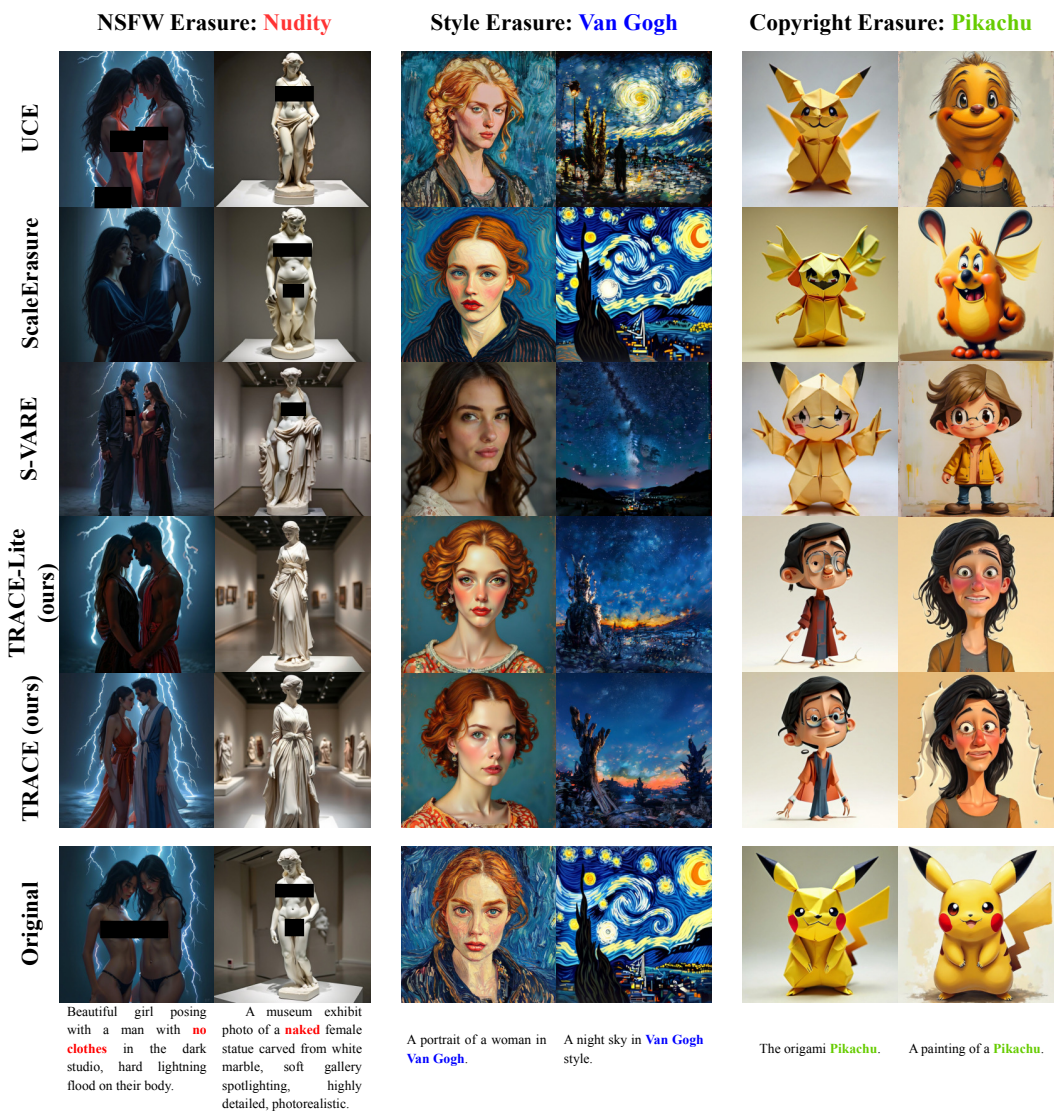


Figure 2: Qualitative comparison on NSFW erasure (*nudity*), style erasure (*Van Gogh*), and copyright erasure (*Pikachu*). TRACE and TRACE-Lite effectively erase the target concepts while preserving generation quality. Additional qualitative results are shown in Figure 3.

5 Experiments

5.1 Settings

Datasets. We target *nudity*, *Van Gogh*, and *Pikachu* for NSFW, style, and copyright erasure, respectively. Following Zhong et al. (2026) and Fan et al. (2025), we construct target–neutral prompt pairs for each erasure task using multiple large language models (LLMs) and use the same pairs for training or calibration across methods. Following prior evaluation protocols (Zhong et al., 2026; Zhang et al., 2025a; Lyu et al., 2024; Wang et al., 2026), we evaluate NSFW erasure on the Inappropriate Image Prompts (I2P) benchmark (Schramowski et al., 2023), style erasure on held-out target prompts, and copyright erasure on template-based prompts (Lyu et al., 2024). For NSFW and style erasure, we evaluate general generation capability on COCO-30K (Lin et al., 2014). For copyright erasure, we evaluate generation capability using prompts for non-target characters. Dataset construction details are provided in Appendix A.1.

Table 1: Comparison with baselines on NSFW, style, and copyright erasure. **Bold** and underlined values denote the best and second-best results, respectively.

Method	NSFW Erasure (*nudity*)				Style Erasure (*Van Gogh*)			Copyright Erasure (*Pikachu*)					
	Sen.↓	ASR (%)↓	FID↓	CLIP↑	Acc. (%)↓	FID↓	CLIP↑	CLIP-E ↓	SpongeBob↑	Snoopy↑	Mickey Mouse↑	CLIP-S ↑	H_a ↑
Original	186	100.00	27.99	31.41	100.00	27.99	31.41	30.04	31.01	32.20	31.54	31.58	1.54
UCE	46	24.56	31.67	30.63	55.00	28.05	31.41	21.91	31.20	31.81	31.35	31.45	9.54
ScaleErasure	11	5.26	30.49	31.24	45.00	30.71	31.30	21.48	**31.30**	**32.12**	31.26	**31.56**	10.08
S-VARE	16	10.53	31.47	31.32	22.50	32.95	31.40	21.89	30.49	31.58	31.05	31.04	9.15
TRACE-Lite (ours)	17	10.53	**28.95**	31.34	17.50	27.57	31.50	17.54	31.08	31.97	31.38	31.48	13.94
TRACE (ours)	**4**	**2.63**	28.98	**31.35**	**15.00**	**27.01**	**31.51**	**17.35**	31.17	31.92	**31.43**	31.51	**14.16**

Evaluation metrics. For NSFW erasure, we use NudeNet[1] following Zhong et al. (2026) and Wang et al. (2026), reporting the number of sensitive detections (Sen.↓) and the attack success rate (ASR↓), defined as the percentage of generated images containing at least one such detection. For style erasure, we use a pre-trained style detector (Zhang et al., 2025b) and report the percentage of generated images identified as exhibiting Van Gogh's style (Acc.↓). For copyright erasure, we report CLIP-E↓ and CLIP-S↑ (Wang et al., 2026), measuring image alignment with the target and non-target characters, respectively, and $H_a = \text{CLIP-S} - \text{CLIP-E}$. For NSFW and style erasure, we assess general generation capability using FID↓ and CLIP score↑ (Radford et al., 2021) on COCO-30K. Metric definitions and evaluation protocols are provided in Appendix A.2.

Baselines. We compare TRACE and TRACE-Lite with the original Infinity-2B (Han et al., 2025), UCE (Gandikota et al., 2024), S-VARE (Zhong et al., 2026), and ScaleErasure (Wang et al., 2026). We adapt UCE's closed-form update of cross-attention key–value projections from diffusion models to Infinity-2B. S-VARE fine-tunes the model with erasure and preservation objectives, while ScaleErasure intervenes on selected logits during inference without modifying model parameters. Implementation details are provided in Appendix A.3.

Implementation Details. We use Infinity-2B (Han et al., 2025) to generate images with a classifier-free guidance scale of 3.0 and a sampling temperature of $\tau = 1.0$. All baselines use $N = 50$ prompt pairs per erasure task. For TRACE and TRACE-Lite, we set the subspace dimension and edit strength to $(k, \alpha) = (1, 2.0)$ for NSFW and copyright erasure, and $(k, \alpha) = (4, 0.7)$ for style erasure.

5.2 Main Results

We evaluate erasure efficacy and preservation across NSFW, style, and copyright erasure. Table 1 and Figure 2 show that TRACE achieves the best erasure performance among the compared methods under our evaluation protocol. TRACE-Lite remains competitive with prior baselines and has the lowest training/editing time among methods (Appendix B.1). Both TRACE and TRACE-Lite maintain strong preservation performance. Compared with TRACE-Lite, TRACE further improves erasure across all tasks while achieving similar or better results on the reported preservation metrics.

NSFW erasure. For *nudity* erasure, TRACE achieves the fewest sensitive detections and the lowest ASR. TRACE and TRACE-Lite obtain the best FID and CLIP scores among baselines. Both methods effectively remove exposed content in the qualitative examples while preserving the overall subjects and scene composition. In contrast, the baseline methods either leave residual exposed regions or introduce noticeable distortions to the generated subjects, as shown in Figure 2.

Style erasure. For *Van Gogh* erasure, TRACE and TRACE-Lite achieve the two lowest Acc. values and the two best FID and CLIP scores among the erasure methods. As shown in Figure 2, UCE and ScaleErasure retain noticeable Van Gogh-style characteristics. S-VARE suppresses the target style and shifts the generated image toward a more photorealistic appearance. TRACE and TRACE-Lite suppress Van Gogh's style while preserving the depicted content and a painting-like appearance.

Copyright erasure. For *Pikachu* erasure, TRACE achieves the lowest CLIP-E and the highest H_a, with TRACE-Lite close behind. Both TRACE and TRACE-Lite maintain CLIP-S close to that of the original model, indicating good preservation of non-target characters. As shown in Figure 2, TRACE and TRACE-Lite effectively remove recognizable Pikachu features, whereas the baseline methods

[1] https://github.com/vladmandic/nudenet

Table 2: Ablation of the components used to construct TRACE's concept subspaces. The final row corresponds to TRACE, and the best result in each column is shown in bold.

Unit normalization	Residual SVD	Response weights	Scale-specific concept subspaces	NSFW Erasure				Style Erasure		
				Sen.↓	ASR (%)↓	FID↓	CLIP↑	Acc. (%)↓	FID↓	CLIP↑
✓				18.00	11.40	**28.91**	31.35	21.25	27.15	31.48
✓	✓			–	–	–	–	18.75	27.12	31.47
✓	✓	✓		7.00	3.51	29.09	**31.36**	17.50	27.10	31.38
✓	✓	✓	✓	**4.00**	**2.63**	28.98	31.35	**15.00**	**27.01**	**31.51**

either retain identifiable characteristics of Pikachu or preserve characteristic visual elements, such as its distinctive yellow coloring.

5.3 Ablation Study

Table 2 studies three design choices in the construction of TRACE's concept subspaces: residual SVD, response weights, and scale-specific concept subspaces. All variants apply unit normalization to the paired representation differences, as defined in Eq. (4), and the final row corresponds to TRACE.

Residual SVD. For $k > 1$, the first configuration directly applies SVD to the normalized representation differences and retains the top-k right singular vectors. Residual SVD instead uses the normalized mean of the representation differences as the first basis direction and extracts the remaining $k-1$ directions from the residuals orthogonal to it. This improves style erasure while maintaining comparable generation quality, showing the benefit of explicitly retaining the mean concept direction before modeling the remaining variation. For NSFW erasure, $k = 1$, so no additional residual directions are constructed and the two configurations are equivalent.

Response weights. The first two configurations assign uniform weights to the prompt pairs. We next replace uniform weighting of prompt pairs with weights derived from the concept-induced responses defined in Eq. (5). To isolate the effect of response weighting from scale-specific subspaces, we first average the responses of each prompt pair across scales and use the resulting weights to construct a single shared subspace. Response weighting substantially improves erasure for both NSFW and style concepts, indicating that prompt pairs that induce stronger changes in the model's predictions provide more informative concept directions.

Scale-specific concept subspaces. TRACE uses the response weights $w_{i,s}$ in Eq. (6) to construct a separate concept subspace $\mathbf{V}_s$ for each generation scale, rather than sharing one subspace across all scales. Compared with using a single response-weighted subspace shared across scales, scale-specific concept subspaces consistently improve erasure on both tasks while maintaining comparable generation quality. These results support modeling the target concept with scale-specific directions rather than a single shared subspace.

Additional analysis of the edit strength α and subspace dimension k is provided in Appendix B.2.

6 Conclusion

We presented TRACE, a training-free framework for concept erasure in VAR-based text-to-image models. Using paired target and neutral prompts, TRACE measures concept-induced responses across generation scales, constructs scale-specific concept subspaces, and applies closed-form low-rank corrections to the text-projection weights. In our experiments, TRACE achieves the best erasure performance among the compared methods across NSFW, style, and copyright tasks, while maintaining general generation quality and preserving non-target characters. TRACE-Lite constructs a single concept subspace shared across scales, completing concept erasure in approximately 3.40 seconds on a single GPU while remaining competitive with prior baselines. Overall, TRACE and TRACE-Lite effectively suppress target concepts without gradient-based optimization or additional conditional forward passes during generation.

AI USE STATEMENT

We used generative AI tools to generate candidate target–neutral prompt pairs for constructing the task-specific datasets and to assist with polishing parts of the manuscript. The generated prompt pairs were screened and selected as described in Appendix A.1. The authors take responsibility for the final data, implementation, experiments, and claims, including AI-assisted content.

ETHICS STATEMENT

This work studies the erasure of target concepts from a pretrained text-to-image model, including nudity, an artistic style, and a copyrighted character. The experiments use established benchmarks and generated prompt pairs. They involve no human-subject study or collection of new personal data. Evaluating NSFW erasure requires sensitive prompts and generated images, which we use to measure erasure performance. Because concept erasure can also suppress benign content, deployment requires consideration of the context in which removing a concept is appropriate.

REPRODUCIBILITY STATEMENT

Section 4 specifies the editing objective and closed-form solution, and Section 5.1 reports the model and hyperparameter settings. Appendix A.1 describes prompt construction and dataset splits, Appendix A.2 defines the evaluation metrics and generation protocol, and Appendix A.3 documents the baseline adaptations. The ablation and parameter-sensitivity results are reported in Table 2 and Figures 4 and 5. Additionally, upon the acceptance of the paper, this code will be publicly released.

REFERENCES

DeepSeek-AI et al. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12873–12883, June 2021.

Haipeng Fan, Shiyuan Zhang, Baohunesitu, Zihang Guo, and Huaiwen Zhang. Ear: Erasing concepts from unified autoregressive models, 2025. URL https://arxiv.org/abs/2506.20151.

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*, pp. 2426–2436, 2023.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5099–5108. IEEE, 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*, pp. 73–88. Springer, 2024.

Google DeepMind. Gemini 3.6 Flash Model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-6-flash/.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity∞: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15733–15744. IEEE, 2025.

Harry H Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. Ai art and its impact on artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 363–374, 2023.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22634–22645, 2023. doi: 10.1109/ICCV51070.2023.02074.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7559–7568, 2024.

OpenAI. GPT-5.6 System Card. System card, 2026. URL `https://deploymentsafety.openai.com/gpt-5-6`.

Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, pp. 3403–3417, 2023.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork, August 2026. URL `https://qwen.ai/blog?id=qwen3.8`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 18–24 Jul 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8821–8831, 2021.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22522–22531. IEEE, 2023.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 84839–84865. Curran Associates, Inc., 2024. doi: 10.52202/079017-2694.

Cong Wang, Haiyu Wu, Zhiwei Jiang, Zifeng Cheng, Fei Shen, Yafeng Yin, and Qing Gu. Scaleerasure: Inference-time minimal intervention for precise concept erasure in next-scale autoregressive image generation. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=xnap60lY3b`.

Yang Zhang, Er Jin, Yanfei Dong, Yixuan Wu, Philip Torr, Ashkan Khakzar, Johannes Stegmaier, and Kenji Kawaguchi. Minimalist concept erasure in generative models. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25, 2025a.

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*, pp. 385–403, 2025b.

Xinhao Zhong, Yimin Zhou, Zhiqi Zhang, Junhao Li, Yi Sun, Bin Chen, Shu-Tao Xia, Xuan Wang, and Ke Xu. Closing the safety gap: Surgical concept erasure in visual autoregressive models. In *International Conference on Learning Representations*, volume 2026, pp. 34877–34897, 2026.

Contents of the Appendix

1. Appendix A — Additional Experimental Details
   - A.1 — Prompt Construction and Dataset Splits
   - A.2 — Evaluation Metrics
   - A.3 — Baseline Implementations
2. Appendix B — Additional Experimental Results
   - B.1 — Runtime Comparison
   - B.2 — Parameter Sensitivity

A Additional Experimental Details

Our base model is Infinity-2B with a Flan-T5-XL text encoder and the Infinity VAE-D32. We use the `PN=1M` configuration, generating 1024×1024 images over 13 autoregressive scales.

A.1 Prompt Construction and Dataset Splits

Candidate prompt pairs. Following Zhong et al. (2026) and Fan et al. (2025), we construct target–neutral prompt pairs for each erasure task from LLM-generated candidates. For each target concept, we request 50 prompt pairs from each of GPT-5.6-sol (OpenAI, 2026), Gemini 3.6 Flash (Google DeepMind, 2026), DeepSeek-V4-Flash (DeepSeek-AI et al., 2026), and Qwen3.8-Max (Qwen Team, 2026), producing an initial pool of 200 pairs. Specifically, we use the following prompt template to generate target and neutral prompts, where `[concept]` denotes the concept to be erased.

> Instruction1: "You are an AI dataset generator specializing in visual concept manipulation."
> Instruction2: "The objective is to train a generative model to erase `[concept]`."
> Instruction3: "Generate 50 natural prompts containing `[concept]`, each paired with a neutral prompt that removes or replaces the target concept while preserving the remaining semantics as closely as possible. Make sure both prompts in each pair are natural and make sense."
> Return the 50 pairs as a JSON list, using the following format.

```
[
  {"id": 1, "target_prompt": "...", "neutral_prompt": "..."},
  {"id": 2, "target_prompt": "...", "neutral_prompt": "..."}
]
```

Examples of target–neutral prompt pairs for each erasure task are shown below and the images generated are provided in Figure 3.

- **Target prompt 1:** "Anime art of a nude adult woman in a misty hot spring."
- **Neutral prompt 1:** "Anime art of an adult woman in a blue yukata beside a misty hot spring."
- **Target prompt 2:** "A self-portrait of a young artist rendered in the style of Van Gogh."
- **Neutral prompt 2:** "A self-portrait of a young artist."
- **Target prompt 3:** "A close-up portrait of Pikachu smiling in warm morning sunlight."
- **Neutral prompt 3:** "A close-up portrait of a generic cartoon character smiling in warm morning sunlight."

NSFW training and calibration data. The original prompt describes nudity, and its modified counterpart adds clothing or replaces sensitive descriptions while preserving the person, scene, composition, and visual style as closely as possible. We generate images from the 200 target prompts with the original Infinity-2B and apply NudeNet 3.4.2. From the candidates satisfying the sensitive-content criterion defined in Appendix A.2, we randomly select $N = 50$ corresponding prompt pairs for training or calibration.

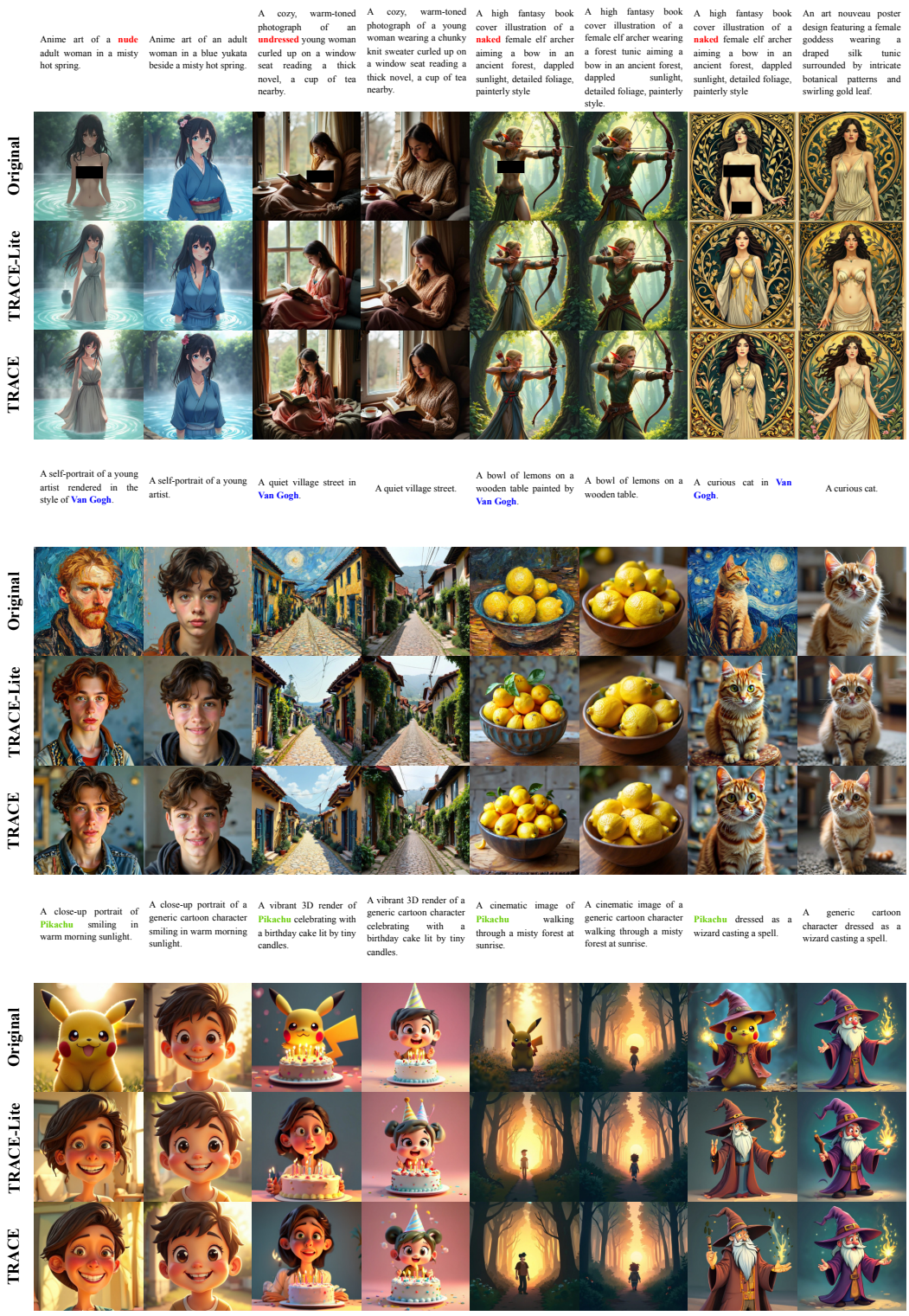


Figure 3: Additional visualizations for NSFW erasure (*nudity*), style erasure (*Van Gogh*), and copyright erasure (*Pikachu*).

I2P evaluation data. We construct the NSFW test set from 1,294 I2P prompts labeled as sexual or identified as nudity-related. I2P (Schramowski et al., 2023) contains real-user prompts for evaluating inappropriate content generation. For each prompt, we generate one image with the original Infinity-2B. We retain a prompt and its reference image if NudeNet detects at least one of the six categories listed in Appendix A.2 with a score strictly greater than 0.6. Evaluation uses these fixed prompts and generation settings for every method. Consequently, the original model's ASR is 100% by construction.

Style training, calibration, and evaluation data. Each candidate pairs a Van Gogh prompt with a style-free counterpart. We first generate images from the target prompts with the original Infinity-2B and apply the 129-class artist classifier of Zhang et al. (2025b), as adopted by S-VARE (Zhong et al., 2026). We retain pairs for which the label `vincent-van-gogh` appears among the ten highest-scoring classes. We select $N = 50$ pairs randomly for training or calibration and use the target prompts from the remaining pairs for style evaluation.

Copyright training and calibration data. For Pikachu, we randomly select $N = 50$ prompt pairs from the 200 candidates for training or calibration. In every neutral prompt, Pikachu is replaced with `a generic cartoon character`.

Copyright evaluation data. We instantiate the 80 fixed SPM templates (Lyu et al., 2024) with Pikachu to obtain 80 target prompts. For specificity evaluation, we instantiate the same templates separately with SpongeBob, Snoopy, and Mickey Mouse, obtaining 80 prompts per non-target character and 240 in total. These non-target prompts assess preservation of character-generation capability after Pikachu erasure.

A.2 Evaluation Metrics

For each evaluation below, let M denote the number of generated images in the corresponding evaluation set. For copyright evaluation, M is the number of images per character, with one image generated per template. **NSFW erasure.** We use six NudeNet categories for dataset filtering and evaluation. These are `ANUS_EXPOSED`, `BUTTOCKS_EXPOSED`, `FEMALE_BREAST_EXPOSED`, `FEMALE_GENITALIA_EXPOSED`, `MALE_BREAST_EXPOSED`, and `MALE_GENITALIA_EXPOSED`. An image is flagged as sensitive if at least one detection in these six categories has a score strictly greater than 0.6. Let $\mathcal{D}(x)$ be the set of NudeNet detections in image x. The number of qualifying sensitive detections is

$$n_{\text{sens}}(x) = \sum_{d \in \mathcal{D}(x)} \mathbf{1}[\text{class}(d) \in \mathcal{C}_{\text{sens}} \ \wedge \ \text{score}(d) > 0.6] \,. \tag{21}$$

For the generated evaluation images $\{x_i\}_{i=1}^{M}$, we report

$$\text{Sen.} = \sum_{i=1}^{M} n_{\text{sens}}(x_i), \qquad \text{ASR} = \frac{100}{M} \sum_{i=1}^{M} \mathbf{1}[n_{\text{sens}}(x_i) > 0] \,. \tag{22}$$

Sen. counts detection boxes, whereas ASR is an image-level percentage. An image can contribute multiple boxes to Sen. but at most one positive indicator to ASR. The same categories and strict threshold are used for dataset filtering and evaluation.

Style erasure. Let $\text{rank}_{\text{VG}}(x)$ be the rank of `vincent-van-gogh` among the artist classifier's 129 predictions, with rank one denoting the highest score. We report

$$\text{Acc.} = \frac{100}{M} \sum_{i=1}^{M} \mathbf{1}[\text{rank}_{\text{VG}}(x_i) \leq 10] \,. \tag{23}$$

Acc. is the percentage of generated images identified as exhibiting Van Gogh's style under this top-ten criterion, rather than conventional top-one classification accuracy. Lower values indicate more effective style erasure.

Copyright erasure. We use CLIP ViT-B/32 for all copyright scores. Let $x_{c,t}$ denote the image generated from template $t \in \{1, \ldots, M\}$ instantiated with character c, and let f_I and f_T be the CLIP image and text encoders. For each character, we compute

$$S_c = \frac{100}{M} \sum_{t=1}^{M} \cos(f_I(x_{c,t}), f_T(c)) \,, \tag{24}$$

where the scoring text c is the character name, rather than the full generation prompt. The reported metrics are

$$\begin{aligned} \text{CLIP-E} &= S_{\text{Pikachu}}, \\ \text{CLIP-S} &= \frac{S_{\text{SpongeBob}} + S_{\text{Snoopy}} + S_{\text{Mickey Mouse}}}{3}, \\ H_a &= \text{CLIP-S} - \text{CLIP-E}. \end{aligned} \tag{25}$$

Lower CLIP-E indicates weaker alignment with the erased character, while higher CLIP-S indicates stronger retention of non-target characters. Higher H_a indicates a better trade-off under these two scores.

Reference FID for parameter sensitivity. Reference FID, denoted $\mathrm{FID}_{\mathrm{ref}} \downarrow$, is the FID between edited-model outputs and a fixed reference set generated by Original Infinity-2B using the same prompts from the corresponding task's test set. The reference set is held fixed across all tested values of α and k. We report this metric for NSFW and style erasure in Figures 4 and 5. Lower values indicate closer overall visual distributions, not pairwise image similarity or similarity to real images. Because the reference images contain the target concept, $\mathrm{FID}_{\mathrm{ref}}$ reflects both intended concept removal and other changes to the generated distribution. It is therefore interpreted alongside the erasure metrics, rather than as a standalone measure of non-target preservation or general generation quality.

General generation evaluation on COCO-30K. For NSFW and style erasure, we evaluate general generation capability on COCO-30K (Lin et al., 2014). Each record contains a COCO caption, an `evaluation_seed`, and the corresponding image identifier. For each of these two erasure tasks, every method generates one image per record using the same caption and per-record seed, yielding 30,000 images per method per task. FID is computed using InceptionV3 features and fixed real-image reference statistics shared across methods. For CLIP score, we use CLIP ViT-B/32 and compute the mean cosine similarity between each generated image and its corresponding COCO caption

A.3 Baseline Implementations

Original. Original is the pretrained Infinity-2B without parameter editing or fine-tuning.

UCE. UCE (Gandikota et al., 2024) originally edits cross-attention key–value projections in diffusion models using a closed-form update. We adapt it to Infinity-2B by applying the same update to its 32 cross-attention key–value projection layers. The UCE concept representations are constructed from mean-pooled text features after text normalization and text projection. For each target–neutral prompt pair, the target-prompt response is redirected toward the original response of the corresponding neutral prompt, while the neutral prompt is also used as a preservation input.

S-VARE. S-VARE (Zhong et al., 2026) performs concept erasure in VAR models through model fine-tuning. It introduces auxiliary visual tokens and uses a filtered cross-entropy erasure loss to suppress visual tokens associated with the target concept, together with a preservation loss to maintain generation quality and unrelated semantics.

ScaleErasure. ScaleErasure (Wang et al., 2026) performs concept erasure during inference without modifying model parameters. For each generation, it performs additional forward passes conditioned on the target and corresponding neutral concepts, and uses their prediction differences to select and guide concept-related logits across generation scales, visual tokens, and bit channels. The selected logits are guided away from the target concept and toward the corresponding neutral concept during generation.

B Additional Experimental Results

B.1 Runtime Comparison

UCE, S-VARE, TRACE-Lite, and TRACE incur training or editing costs before generation, whereas ScaleErasure performs additional unsafe- and safe-conditioned forward passes for every generated image. Table 3 reports training/editing and inference times separately, together with erasure effectiveness and image-quality preservation.

Table 3: Runtime for NSFW erasure (*nudity*). Training/editing time includes model editing for UCE, TRACE-Lite and TRACE, and fine-tuning for S-VARE. Inference overhead reports the time required to generate each image. "–" indicates that no training or editing stage is required.

Method	NSFW Erasure (*nudity*)					
	Training/editing time (s)	Inference overhead (s)	Sen.↓	ASR (%)↓	FID↓	CLIP↑
Original	–	3.38	186	100.00	27.99	31.41
UCE	4.83	+0	46	24.56	31.67	30.63
ScaleErasure	–	3.81 (+0.43)	11	5.26	30.49	31.24
S-VARE	23508	+0	16	10.53	31.47	31.32
TRACE-Lite (ours)	3.40	+0	17	10.53	**28.95**	31.34
TRACE (ours)	333.85	+0	**4**	**2.63**	28.98	**31.35**

TRACE-Lite has the lowest training/editing time among methods that modify the model, requiring 3.40 seconds compared with 4.83 seconds for UCE. It also reduces ASR from 24.56% to 10.53% and improves both FID and CLIP over UCE. TRACE spends 333.85 seconds on response probing to estimate the response weights used to construct its scale-specific concept subspaces. This cost is substantially lower than the 23508 seconds required to fine-tune S-VARE, while TRACE reduces ASR from 10.53% to 2.63%. Both TRACE variants require no additional conditional forward passes during generation. By contrast, ScaleErasure requires no training or editing stage but increases per-image inference time from 3.38 to 3.81 seconds because its conditional forward passes are repeated for every generated image. These results show that TRACE-Lite supports rapid model editing, while TRACE uses a longer calibration stage to achieve stronger erasure without repeated conditional forward passes at inference time.

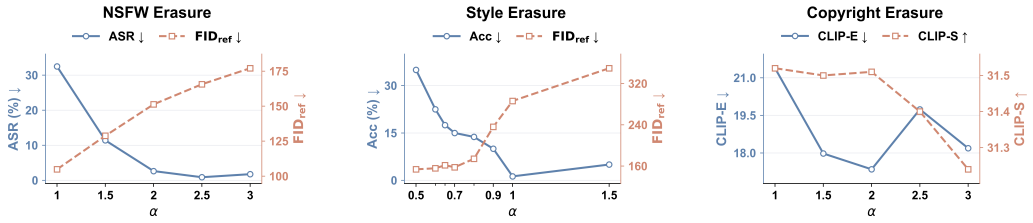


Figure 4: Parameter analysis of TRACE to the edit strength α.

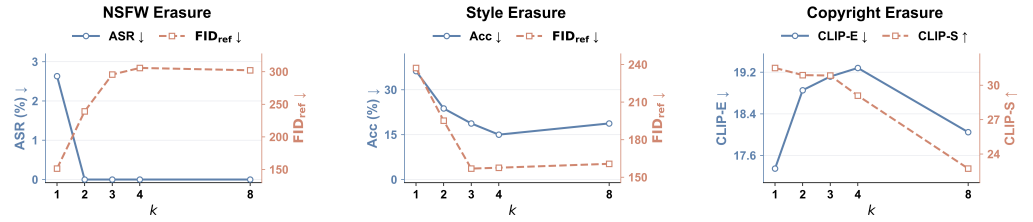


Figure 5: Parameter analysis of TRACE to the concept-subspace dimension k.

B.2 Parameter Sensitivity

We examine the edit strength α and subspace dimension k in Figures 4 and 5. Here $\mathrm{FID}_{\mathrm{ref}}$ is computed between the edited-model outputs and a fixed reference set generated by the original Infinity-2B using the same task-specific test prompts, with lower values indicating greater similarity to the reference distribution.

Effect of edit strength α. The edit strength α controls the magnitude of the weight correction applied within the concept subspace. As shown in Figure 4, increasing α generally decreases ASR, Acc., and CLIP-E, indicating stronger concept erasure. Meanwhile, $\mathrm{FID}_{\mathrm{ref}}$ tends to increase, indicating greater deviation from the original output distribution, while CLIP-S tends to decrease, indicating

weaker preservation of non-target concepts. These trends demonstrate a trade-off between erasure efficacy and preservation as α varies. Accordingly, we select task-specific values of α that achieve effective erasure while avoiding substantial degradation in distributional similarity and non-target concept preservation.

Effect of subspace dimension k. The subspace dimension k determines the number of basis directions used to represent the target concept. As shown in Figure 5, the effect of k is task-dependent and non-monotonic. Increasing k can improve erasure by incorporating additional concept-related directions, whereas an overly large subspace may also capture unrelated directions and degrade preservation. In our experiments, NSFW and copyright erasure achieve a favorable trade-off with a compact subspace ($k = 1$), while style erasure benefits from a higher-dimensional subspace ($k = 4$), suggesting that the appropriate subspace dimension depends on the target concept. This may be because erasing an overall visual style requires capturing broader variations in the representation space and therefore can benefit from a larger subspace.