

# Now You See It, Now You Don’t - Instant Concept Erasure for Safe Text-to-Image and Video Generation

Shristi Das Biswas, Arani Roy, Kaushik Roy  
Purdue University

{sdasbisw, roy173, kaushik}@purdue.edu

## Abstract

*Robust concept removal for text-to-image (T2I) and text-to-video (T2V) models is essential for their safe deployment. Existing methods, however, suffer from costly re-training, inference overhead, or vulnerability to adversarial attacks. Crucially, they rarely model the latent semantic overlap between the target erase concept and surrounding content – causing collateral damage post-erasure – and even fewer methods work reliably across both T2I and T2V domains. We introduce Instant Concept Erasure (ICE), a training-free, modality-agnostic, one-shot weight modification approach that achieves precise, persistent unlearning with zero overhead. ICE defines erase and preserve subspaces using anisotropic energy-weighted scaling, then explicitly regularizes against their intersection using a unique, closed-form overlap projector. We pose a convex and Lipschitz-bounded Spectral Unlearning Objective, balancing erasure fidelity and intersection preservation, that admits a stable and unique analytical solution. This solution defines a dissociation operator that is translated to the model’s text-conditioning layers, making the edit permanent and runtime-free. Across targeted removals of artistic styles, objects, identities, and explicit content, ICE efficiently achieves strong erasure with improved robustness to red-teaming, all while causing only minimal degradation of original generative abilities in both T2I and T2V models.*

## 1. Introduction

The generative AI revolution, powered by text-to-image (T2I) models [1–5] and text-to-video (T2V) models [6–8], offers immense creative power but also significant risks. Trained on large-scale, uncurated internet datasets [9], these models can be easily prompted to generate harmful, copyrighted, or undesired content [10], including explicit material [11, 12], copyrighted artistic styles [13–16], or the identities of specific individuals [17, 18]. This creates an urgent and critical need for robust, efficient, and precise mechanisms to control model outputs by erasing undesired con-

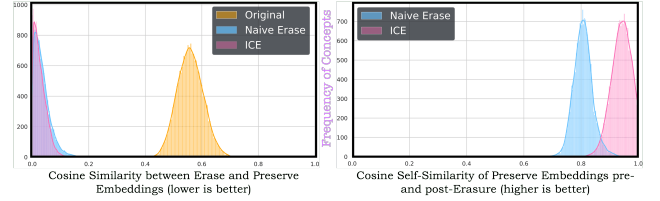


Figure 1. ICE effectively isolates and removes target erase ( $e$ ) concepts while minimizing impact on a preserved ( $p$ ) set by explicitly addressing their semantic overlap. **(L)** Similarity between  $e$  and  $p$  embeddings; lower implies less overlap: ICE significantly reduces the Original embedding similarity (yellow), outperforming Naive Erasure (blue). **(R)** Self-similarity measuring preservation of  $p$  embeddings after erasure. Higher means lesser collateral damage. ICE’s preservation of original semantics is superior to Naive Erase.

cepts [19–21], thereby enabling a safer and more responsible deployment of generative models.

Current approaches to this problem fall into three main categories, each with their own significant drawbacks. Inference-time methods apply safety filters or guidance during generation [12, 22, 23]. While easy to implement, they introduce significant computational overhead during each inference-run and are often brittle, easily bypassed by simple adversarial prompts [24, 25]. On the other hand, training-based methods retrain or finetune the model to ‘forget’ a concept [26–31]. This approach is prohibitively expensive, data-intensive, and often leads to catastrophic forgetting, degrading the model’s general generative quality. A more promising third category, training-free weight modification, directly edits the model’s parameters in a single shot [32–34]. These methods are fast and incur no inference overhead. However, they fail to explicitly address the fundamental challenge of *semantic overlap*. In the high-dimensional latent space, the subspace for an erase concept (e.g.,  $\mathcal{S}_e$  for “Van Gogh”) is inherently non-orthogonal to related preserve concepts (e.g.,  $\mathcal{S}_p$  for “painting” or an unconditional “ ” [35]). Naive erasure via simple orthogonal projection [22, 34, 36, 37] (as seen in Fig. 1 (R)) inevitably inflicts collateral damage on the shared semantic region,  $\mathcal{S}_e \cap \mathcal{S}_p$ , degrading the quality of the related non-

targeted concepts post-erasure (also viewed through the lens of set-theoretic difference in App. Sec E.3). This failure to explicitly model and preserve the overlap is the primary reason for their trade-offs in fidelity and robustness. Furthermore, most prior work focuses exclusively on T2I models, neglecting the growing need for modality-agnostic safety methods that also lend themselves to safe T2V generation.

To overcome these limitations, we propose **Instant Concept Erasure (ICE)** – a unified, training-free framework that performs precise, one-shot concept removal through a closed-form, geometry-aware formulation. ICE operates directly in the semantic conditioning space of generative diffusion models, identifying and isolating the latent subspace responsible for an undesired concept while explicitly preserving overlapping semantics shared with other safe or general concepts. Unlike prior methods that rely on iterative finetuning or run-time masking, ICE derives a formal Spectral Unlearning Objective whose global minimum admits a unique, analytic solution to this task, enabling an instantaneous and permanent edit to model parameters.

Our key insight is that effective unlearning requires explicitly characterizing and protecting the *intersection* between erase and preserve subspaces, rather than assuming they are orthogonal or disjoint. By leveraging energy-weighted bases [38] derived from the target embeddings’ singular spectrum, ICE builds anisotropic subspace operators for both the erase concept ( $e$ ) and its complementary preserve set ( $p$ ), and computes their intersection using a closed-form overlap projector. This formulation yields an interpretable and stable dissociation operator that directly specifies the optimal forgetting direction, which we then translate into a permanent one-shot weight update at a given generative model’s earliest text-conditioning interface (e.g., the key-value projections in UNet-based models or the text-projection layers in diffusion transformers).

ICE is both *modality-agnostic* and *efficient*: it applies identically to T2I and T2V models in only 2 seconds, requiring no retraining or additional inference-time cost. We demonstrate its broad applicability across four major unlearning axes – *artistic styles*, *object categories*, *identities*, and *explicit content* – showing strong erasure efficacy with minimal damage to the model’s base generation quality. Further, ICE substantially enhances resistance to adversarial and red-teaming attacks, outperforming prior training-free and finetuning-based unlearning methods. In summary, ICE provides a mathematically grounded and computationally lightweight solution to safe concept removal, achieving reliable, fast, and generalizable unlearning across generative modalities without sacrificing fidelity or robustness.

## 2. Related Works

**Safe T2I Generation.** Current methods for removing undesired concepts T2I models can be classified into

three paradigms, each with distinct trade-offs. The first, *inference-time control*, applies post-hoc interventions such as safety checkers [39], classifier-free guidance during generation [12, 40] or filtering the embeddings away from identified unsafe subspaces and using adaptive token-wise shifts to navigate similarity between run-time prompt tokens and target concept tokens [22, 23]. While easy to implement, they require independent safeguarding operations for every new prompt, hindering inference-time efficiency, and are easily circumvented in open-source models. The second paradigm, *training-based interventions*, involves modifying the model through retraining or finetuning [30, 41–43]. This includes techniques like training on filtered datasets, using negative guidance, self-supervised learning or adversarial training [26, 44–48], minimizing KL divergence between unwanted and alternative safe concepts [27], and employing preference optimization to suppress concepts [49, 50]. On the other hand, partial finetuning approaches adjust specific layers to forget undesired concepts [28, 29, 31]. Although effective to some degree, these methods are computationally expensive, require careful data curation, and may degrade general model performance while still remaining vulnerable to adversarial attacks [51]. The third category consists of *training-free weight modification* techniques, which aim to erase concepts via one-shot updates to model parameters [32–34, 52]. While they avoid the cost of retraining, they often fail to safeguard overlapping semantics with untargeted surrounding concepts, and are easily bypassed by black-box or white-box red-teaming tools [24, 25, 53, 54].

**Safe T2V Generation.** The domain of concept unlearning for T2V generation is still nascent, with few existing works. These pioneering approaches can be broadly categorized. One strategy involves intervention in the embedding space [22] that identifies and removes harmful representations from text embeddings during run-time. A second, transfer-based approach leverages unlearning from the more mature T2I domain [55] to optimize a text encoder using few-shot image-based unlearning and then deploys this finetuned ‘safe’ encoder in a T2V model. A third work involves direct finetuning of the T2V model using negatively-guided velocity prediction [56]. These existing methods rely on inference-time filtering, knowledge transfer, or computationally intensive finetuning. Importantly, they also fail to provide explicit preservation for the untargeted concepts.

Differently, ICE is a unified, modality-agnostic edit for both T2I and T2V, uniquely designed to be intersection-aware. We pose a convex Spectral Unlearning Objective that regularizes the update with a formal, non-iterative characterization of the semantic overlap between importance-weighted erase and preserve subspaces, and prove that it admits a unique closed-form dissociation direction. This operator translates into a permanent weight update, enabling robust, high-fidelity unlearning with zero inference overhead.



### 3. Method

We achieve this through the ICE closed-form, geometrical overlap-aware formulation for one-shot concept erasure. The process is elucidated as follows.

#### 3.1. Problem Formulation in Conditioning Subspaces

We formalize this task by considering a single network layer where an output activation  $\mathbf{o} \in \mathbb{R}^{1 \times V}$  is computed from input activation  $\mathbf{x} \in \mathbb{R}^{1 \times d}$  and a weight matrix  $\mathcal{W} \in \mathbb{R}^{V \times d}$  via the linear transformation  $\mathbf{o} = \mathbf{x}\mathcal{W}^\top$ .

Our objective is to produce a modified output,  $\mathbf{o}^*$ , from which a target erase concept’s semantic contribution is selectively removed without affecting model performance on untargeted content. We postulate that a concept’s semantics are encoded within an euclidean subspace of the embedding space  $\mathbb{R}^d$ . Erasure can thus be achieved by projecting these semantics out of the input embedding. To this end, we seek an Instant Concept Erasure (ICE) operator,  $\mathcal{P}_{ice} \in \mathbb{R}^{d \times d}$ , that isolates the target concept’s precise representation from all neighboring content. Applying this operator yields the erase concept specific embedding  $\mathbf{x}_{ice} = \mathbf{x}\mathcal{P}_{ice}$  (to be removed), which in turn produces the ideal unlearned output:

$$\mathbf{o}^* = (\mathbf{x} - \mathbf{x}_{ice})\mathcal{W}^\top = (\mathbf{x}(\mathcal{I} - \mathcal{P}_{ice}))\mathcal{W}^\top \quad (1)$$

This prevents the model from generating the erased concept any further with high specificity. This operation can henceforth be interpreted either as a dynamic filter on the activation  $\mathbf{x}$  or as a one-time persistent update to the model’s parameters by translating the operator to weight space via an updated weight matrix,

$$\mathcal{W}_{ice}^\top = ((\mathcal{I} - \mathcal{P}_{ice})\mathcal{W}^\top) \quad (2)$$

The central challenge, which the remainder of this paper addresses, is the principled, data-driven construction of this precise unlearning operator  $\mathcal{P}_{ice}$ .

#### 3.2. Formal Characterization of Target Subspaces

The latent semantic subspaces for erasure ( $\mathcal{S}_e$ ) and preservation ( $\mathcal{S}_p$ ) are characterized by forming embedding matrices,  $\mathcal{E}_{e,p} \in \mathbb{R}^{d \times N}$ , from their respective prompt tokens via the model’s frozen text encoder [57–59]. Singular Value Decomposition ( $\mathcal{E} = \mathcal{U}\Sigma\mathcal{V}^\top$ ) yields an orthonormal basis for each subspace from the columns of the left singular bases  $\mathcal{U}$  and a corresponding singular value spectrum  $\Sigma$ .

Conventional subspace methods often assume all basis vectors are equally important, an assumption we find too restrictive for precise unlearning [36]. We observe that the singular values exhibit a non-uniform distribution [38], implying that each basis vector ( $\Sigma^i$ ) contributes unequally to the concept’s representation. Leveraging this, we propose assigning an anisotropic importance weight,  $\lambda^i$ , to each basis vector  $\mathcal{U}^i$ . The importance is computed as follows:

$$\lambda^i = \frac{2\Sigma^i}{\Sigma^i + \max_{j=1..k}(\Sigma^j)}, \text{ where } \lambda^i \in [0, 1] \quad (3)$$

This fixed weighting profile assigns maximal salience ( $\lambda = 1$ ) to the principal component while other bases are given importance ( $< 1$ ) relative to this maximum. These importance scores,  $\lambda_e$  and  $\lambda_p$ , form diagonal matrices,  $\Lambda_e$  and  $\Lambda_p$ , which define the scaled spectral conditioning operators:

$$\mathcal{P}_e = \mathcal{U}_e\Lambda_e\mathcal{U}_e^\top \quad \text{and} \quad \mathcal{P}_p = \mathcal{U}_p\Lambda_p\mathcal{U}_p^\top \quad (4)$$

This characterization creates contractive attenuators that emphasize high-energy, concept-defining directions while suppressing low-energy spaces that are generic and shared across neighboring concepts.

#### 3.3. The Challenge of Semantic Overlap and Subspace Intersection

The core limitation in existing unlearning literature is the lack of dissociation of concepts targeted for erasure ( $e$ ) from those intended for preservation ( $p$ ). As discussed, this primary challenge arises from the geometric relationship between their corresponding latent subspaces,  $\mathcal{S}_e$  and  $\mathcal{S}_p$ . We further take into consideration the practical challenge of defining an ideal set  $p$ , and hence design it to encompass the model’s generic, average knowledge of all other possible concepts. We use the unconditional embedding “ ” as a principled and efficient proxy for this space, as its role as the model’s generative prior was established by [35]. While an ideal, disentangled space would feature embedding orthogonality ( $\mathcal{S}_e \perp \mathcal{S}_p$ ) and hence no semantic overlap, we find that this is not the case in practice even after using anisotropic projectors (Eq. 4) as seen in Fig. 1. Consequently, a principled unlearning method must explicitly characterize and preserve the non-trivial overlap  $\mathcal{S}_e \cap \mathcal{S}_p$ . To this end, we establish that this semantic intersection can be characterized exactly and non-iteratively [60, 61]. We formalize this foundational challenge as:

**Proposition 1.** *The unique overlap projector,  $\mathcal{P}_{e \cap p}$ , onto the intersection of target latent subspaces  $\mathcal{S}_e$  and  $\mathcal{S}_p$  is given in closed-form by:*

$$\mathcal{P}_{e \cap p} = 2\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger\mathcal{P}_p \quad (5)$$

where  $\dagger$  denotes the Moore-Penrose pseudoinverse.

The detailed proof is presented in App. Sec. A. This provides a tractable formulation of the overlap projector for our precise unlearning objective, allowing for a non-iterative solution constrained against the subspace of semantic overlap.

#### 3.4. The Spectral Unlearning Objective

With the operators for the target,  $\mathcal{P}_e, p$ , and the intersection subspaces  $\mathcal{P}_{e \cap p}$  established, we isolate the erase-specific embedding  $\mathbf{x}_{ice} \in \mathbb{R}^d$  – excluding shared components with the preserve subspace – via a convex unlearning objective that balances erasure fidelity and overlapping semantics preservation. Our objective  $\mathcal{L}(\mathbf{x}_{ice})$  is:

$$\min_{\mathbf{x}_{ice}} \mathcal{L}(\mathbf{x}_{ice}) \triangleq \min_{\mathbf{x}_{ice}} \underbrace{\|\mathbf{x}_{ice} - \mathbf{x}\mathcal{P}_e\|_2^2}_{\text{Fidelity Term}} + \underbrace{\|\mathbf{x}_{ice}\mathcal{P}_{e\cap p}\|_2^2}_{\text{Preservation Term}} \quad (6)$$

The components of this objective serve specific roles:

- **Fidelity Term:** This quadratic loss term enforces the erasure goal. It minimizes the distance between the unlearning direction  $\mathbf{x}_{ice}$  and the projection of the original concept  $\mathbf{x}$  onto the erase subspace. This ensures  $\mathbf{x}_{ice}$  accurately captures the features designated for ablation.
- **Preservation Term:** This term acts as a convex regularizer to prevent collateral damage to surrounding concepts. It penalizes  $\mathbf{x}_{ice}$  from having any component that lies within the semantic intersection, which is precisely characterized by the overlap projector  $\mathcal{P}_{e\cap p}$ .

Together, these make  $\mathbf{x}_{ice}$  effectively isolate target erase concepts  $e$  while minimizing impact on a preserved  $p$  set by explicitly removing their semantic overlap. A key advantage of this formulation is that  $\mathcal{L}(\mathbf{x}_{ice})$  is strongly convex with Lipschitz-continuous gradient, a property we formally prove in App. Sec. B and Sec. C. This guarantees a unique and stable global minimum, allowing us to bypass iterative optimization methods and derive a direct closed-form solution for the optimal unlearning direction, as shown below.

**Proposition 2.** *Given a target concept embedding  $\mathbf{x}$ , the erase operator  $\mathcal{P}_e$ , and the intersection projector  $\mathcal{P}_{e\cap p}$ , the vector  $\mathbf{x}_{ice}$  that minimizes the objective in Eq. 6 is given by the unique, closed-form solution:*

$$\mathbf{x}_{ice} = \mathbf{x}\mathcal{P}_{ice} \text{ where } \mathcal{P}_{ice} = \mathcal{P}_e(\mathcal{I} + \mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T)^{-1} \quad (7)$$

and,  $\mathcal{P}_{ice} \in \mathbb{R}^{d \times d}$ . The first term in the inverse matrix,  $\mathcal{I}$ , can be interpreted as matching the covariance of the large encyclopedia of concept embeddings in the diffusion model’s vocabulary, as inspired by [62], while the second term is the covariance matrix of the semantic intersection. *Proof.* To find the minima of the convex and Lipschitz-bounded loss  $\mathcal{L}(\mathbf{x}_{ice})$ , we compute its gradient with respect to the row vector  $\mathbf{x}_{ice}$  and set it to zero. Using the identity  $\nabla_{\mathcal{X}} \|\mathcal{X}\mathcal{A} - \mathcal{B}\|_2^2 = 2(\mathcal{X}\mathcal{A} - \mathcal{B})\mathcal{A}^T$ , the gradient is:

$$\nabla_{\mathbf{x}_{ice}} \mathcal{L}(\mathbf{x}_{ice}) = 2(\mathbf{x}_{ice} - \mathbf{x}\mathcal{P}_e) + 2(\mathbf{x}_{ice}\mathcal{P}_{e\cap p})(\mathcal{P}_{e\cap p})^T \quad (8)$$

Setting  $\nabla_{\mathbf{x}_{ice}} \mathcal{L} = 0$ , we isolate  $\mathbf{x}_{ice}$

$$\mathbf{x}_{ice} = \mathbf{x}(\mathcal{P}_e)(\mathcal{I} + \mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T)^{-1} \quad (9)$$

$$= \mathbf{x}\mathcal{P}_{ice} \text{ where } \mathcal{P}_{ice} \in \mathbb{R}^{d \times d} \quad (10)$$

We provide more details on the formalized derivation for  $\mathcal{P}_{ice}$  in App. Sec. D. The existence of this analytical solution is the backbone of ICE’s precision and efficiency.

### 3.5. Update Translation for Persistent Unlearning

The analytical solution from ICE provides the optimal unlearning direction,  $\mathbf{x}_{ice} \in \mathbb{R}^{1 \times d}$ . To apply this solution, we translate the  $\mathcal{P}_{ice}$  activation-space vector into a permanent, one-shot modification of the model’s weight space. This

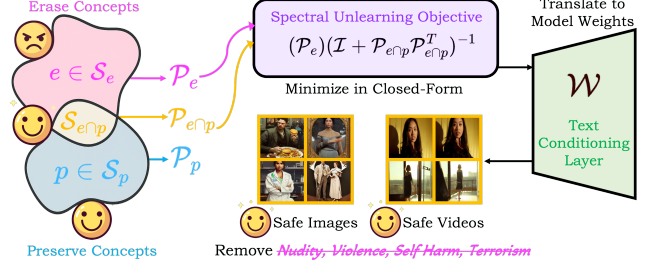


Figure 2. ICE performs one-shot, training-free concept erasure by constructing erase ( $\mathcal{S}_e$ ), preserve ( $\mathcal{S}_p$ ), and intersection ( $\mathcal{S}_{e\cap p}$ ) subspaces. The closed-form operator  $\mathbf{x}(\mathcal{P}_e)(\mathcal{I} + \mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T)^{-1}$  isolates undesired semantics while safeguarding shared content, which is applied as a weight update to the target text-conditioning layer  $\mathcal{W}$  to enable precise, safe, and modality-agnostic unlearning.

procedure ensures the unlearning is both persistent and incurs zero overhead during inference, providing a broadly applicable method for safe T2I and T2V generation.

The generalizability of our method stems from targeting the ‘entry points’ where text conditioning is first injected into the main visual backbone to steer the generative process. This intervention point is functionally consistent across different diffusion architectures. By modifying the model at this initial stage of semantic injection, we can ablate undesirable concepts before they propagate through the generative network, providing a powerful tool for enforcing safety constraints on any text-conditioned model. The primary entry points include:

- For **UNET-based architectures** (e.g., Stable Diffusion), the intervention targets the key ( $\mathcal{W}_k$ ) and value ( $\mathcal{W}_v$ ) projection matrices within each cross-attention (CA) block.
- For many **Diffusion Transformer (DiT) architectures** (e.g., CogVideoX), the intervention targets a dedicated text projection layer that processes text embeddings before they are consumed by the transformer blocks.

The ICE operator is then used to update the weight matrix  $\mathcal{W}_{old}$  of the target entry-point layer(s) in a single shot using Eq. (2). This one-time instant update permanently embeds the unlearning transformation into the T2I/V model’s parameters, resulting in zero additional run-time latency. An overview of the method is illustrated in Fig. 2.

## 4. Evaluation

In this section, we present the results of our method for erasing inappropriate concepts, artistic styles, objects, and identities, as well as resistance to red-teaming attacks. We use StableDiffusion-v1.4 (SD-v1.4) [3] as a primary T2I backbone, following recent work [23, 32, 33]. We further extend evaluations to StableDiffusion-v1.5 [66], StableDiffusion-v2.1 [67] and several community versions of SD, including RealisticVision [68], Dreamshaper [69], and Chilloutmix [70]. Next, we evaluate erasure efficacy for T2V generation by applying to the popular CogVideoX-2B and

| Method          | Weights Mod. | Training -Free | Semantic Overlap Aware | Attack Success Rate ↓ |              |                  |                    |                     | COCO         |             |
|-----------------|--------------|----------------|------------------------|-----------------------|--------------|------------------|--------------------|---------------------|--------------|-------------|
|                 |              |                |                        | I2P [12]              | P4D [53]     | Ring-A-Bell [63] | MMA-Diffusion [54] | UnlearnDiffAtk [24] | FID [64] ↓   | CLIP [65] ↑ |
| SD-v1.4         | -            | -              | -                      | 0.178                 | 0.987        | 0.831            | 0.957              | 0.697               | -            | 31.3        |
| SLD-Medium [12] | ✗            | ✓              | ✗                      | 0.142                 | 0.934        | 0.660            | 0.942              | 0.648               | 31.47        | 31.0        |
| SLD-Strong [12] | ✗            | ✓              | ✗                      | 0.131                 | 0.814        | 0.620            | 0.920              | 0.570               | 40.88        | 29.6        |
| SLD-Max [12]    | ✗            | ✓              | ✗                      | 0.115                 | 0.602        | 0.570            | 0.837              | 0.479               | 50.51        | 28.5        |
| SAFREE [22]     | ✗            | ✓              | ✗                      | 0.272                 | 0.384        | 0.114            | 0.585              | 0.282               | 36.35        | 31.1        |
| AdaVD [23]      | ✗            | ✓              | ✗                      | 0.045                 | <u>0.106</u> | 0.114            | 0.461              | 0.275               | 35.99        | <b>31.3</b> |
| ESD [26]        | ✓            | ✗              | ✗                      | 0.140                 | 0.750        | 0.528            | 0.873              | 0.761               | 52.06        | 30.7        |
| SA [31]         | ✓            | ✗              | ✗                      | 0.062                 | 0.623        | 0.239            | 0.205              | 0.268               | 54.98        | 30.6        |
| CA [27]         | ✓            | ✗              | ✗                      | 0.078                 | 0.639        | 0.376            | 0.855              | 0.866               | 40.99        | <u>31.2</u> |
| MACE [28]       | ✓            | ✗              | ✗                      | <b>0.023</b>          | 0.142        | 0.076            | 0.183              | <b>0.176</b>        | 52.24        | 29.4        |
| SDID [47]       | ✓            | ✗              | ✗                      | 0.270                 | 0.931        | 0.646            | 0.907              | 0.637               | 22.99        | 30.5        |
| CPE [29]        | ✓            | ✗              | ✗                      | 0.046                 | 0.110        | <u>0.012</u>     | 0.186              | <u>0.234</u>        | <u>22.01</u> | 29.7        |
| UCE [33]        | ✓            | ✓              | ✗                      | 0.103                 | 0.667        | 0.331            | 0.867              | 0.430               | 31.25        | <b>31.3</b> |
| RECE [32]       | ✓            | ✓              | ✗                      | 0.064                 | 0.381        | 0.134            | 0.675              | 0.655               | 37.60        | 30.9        |
| CURE [34]       | ✓            | ✓              | ✗                      | 0.061                 | 0.107        | 0.013            | <b>0.169</b>       | 0.281               | -            | -           |
| ICE (Ours)      | ✓            | ✓              | ✓                      | <u>0.043</u>          | <b>0.102</b> | <b>0.011</b>     | <u>0.173</u>       | 0.266               | <b>21.94</b> | <u>31.2</u> |

Table 1. Comparison on Attack Success Rate (ASR). ↓ indicates lower is better, while ↑ shows higher is better. Our method achieves the best performance on the majority of the adversarial benchmarks while maintaining competitive performance on COCO-30k. The COCO scores for CURE are not reported as their implementation is not publicly available. We gray out training-based methods for a fair comparison, while methods in pink require repeated run-time application. Best results are **bolded** and second best underlined.

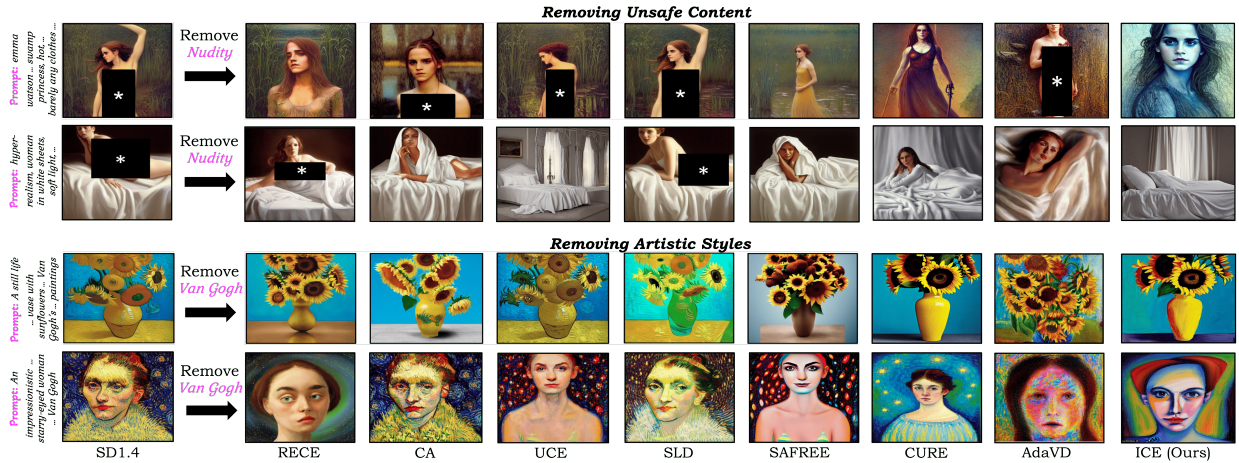


Figure 3. Comparison of unlearning methods on removing target artist styles and NSFW content. ICE suppresses the intended erase concept more effectively than baselines. \* masks any unsafe outputs for display purposes.

CogVideoX-5B models [7]. The preserve set is defined as an unconditional embedding “ ” unless otherwise specified. More details are provided in the Appendix.

#### 4.1. Unsafe Content Erasure

We first evaluate T2I safety using multi-suite red-teaming and report attack-success rate (ASR) alongside COCO [71] utility to demonstrate specificity in preserving normal content post-unlearning. To this end, we test on inappropriate prompts from I2P [12], white-box attacks [24, 53], and black-box attacks [54, 63]. For nudity analysis, we generate one image per prompt and detect unsafe regions with a NudeNet [72] threshold of 0.6, following [32]. As summarized in Tab. 1, ICE – the only method that is simultaneously training-free, a one-shot weight edit, and explicitly overlap-aware – achieves SOTA or second-best ASR across all baselines. In particular, ICE attains 4%, and 31% lower ASR compared to the best-performing counterparts on P4D and

Ring-A-Bell, respectively, highlighting its strong resilience against adversarial attacks. Crucially, ICE delivers superior or comparable safeguarding performance while generating higher-quality images on the COCO-30k dataset, all within a training-free framework (Tab. 1 and App. Fig. 12). Qualitative examples (Fig. 3) draw a similar conclusion, showing precise suppression of unsafe content with minimal collateral damage. Notably, ICE is the only method in Row 1 that preserves Emma Watson’s identity while erasing nudity, underscoring strong preservation of untargeted content.

For T2V, we assess nudity and broader toxicity suppression on three sets: (1) *Gen*, a contextual prompt suite describing nudity in rich detail [23]; (2) *Ring-A-Bell*, comprising short stylized prompts depicting explicit artwork [23, 63]; and (3) *SafeSora* [73], from which we use a 296 toxic prompts across 5-categories, following [22]. For each prompt, we generate 49 frames per model at default resolution and report the NudeNet [72] Nudity Rate. On



| Methods       | SafeSora ↓   |              | Nudity Rate (Gen) ↓ |              | Nudity Rate (Ring-A-Bell) ↓ |             | VBench Object Class ↑ | VBench Subject Consistency ↑ |
|---------------|--------------|--------------|---------------------|--------------|-----------------------------|-------------|-----------------------|------------------------------|
|               | CogX-2B      | CogX-5B      | CogX-2B             | CogX-5B      | CogX-2B                     | CogX-5B     |                       |                              |
| Original      | 64.98        | 79.56        | 57.10               | 61.80        | 30.25                       | 42.50       | 90.75                 | 95.77                        |
| NegPrompt     | 55.73        | 64.20        | 36.00               | 46.35        | 11.75                       | 14.91       | <b>90.64</b>          | 91.62                        |
| SAFREE        | 43.22        | 51.58        | 32.43               | 35.12        | 14.23                       | 10.64       | 53.29                 | <b>93.36</b>                 |
| T2VUnlearning | <u>31.56</u> | <u>39.77</u> | <u>19.73</u>        | <u>16.47</u> | <b>6.97</b>                 | <b>2.74</b> | 85.30                 | 92.09                        |
| ICE (Ours)    | <b>24.01</b> | <b>39.34</b> | <b>11.25</b>        | <b>8.75</b>  | <u>7.84</u>                 | <u>2.98</u> | <u>87.11</u>          | <u>92.38</u>                 |

Table 2. Comparison of safe T2V generation performance across baselines. To assess video generation ability for benign concepts post-unlearning, we evaluate on the VBench benchmark metrics.

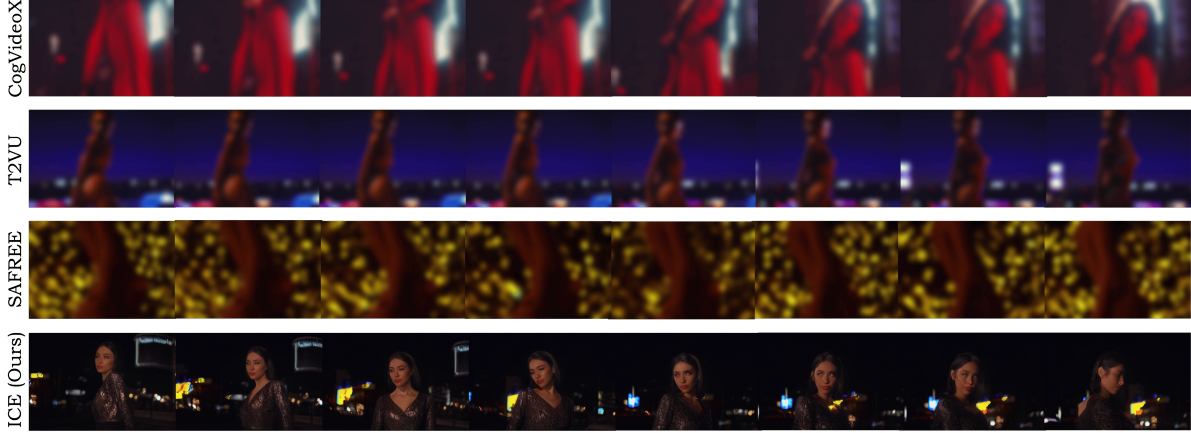


Figure 4. T2V generated examples with CogVideoX-5B on Ring-A-Bell prompts. We manually blurred out generated sexually explicit videos and censored sensitive text prompts for display purposes.

CogVideoX-2B/5B, ICE achieves the lowest SafeSora toxicity and lowest Nudity Rate on Gen by up to 23.9% and 47.7% respectively. Qualitative results show robust nudity suppression with preserved motion and content (Fig. 4), as well as resistance to concepts prompting violence and animal abuse (App. Fig. 10), showing that ICE effectively removes diverse sensitive content while preserving detail and untargted concepts. To analyze the potential impact of nudity erasure on non-nudity concepts, we adopt VBench [74], a widely used video generation benchmark for further evaluation. ICE maintains strong video utility on Object Class and Subject Consistency metrics, as summarized in Tab. 2 and App. Fig. 15. More results in the Appendix.

## 4.2. Artistic Style Erasure

To evaluate the efficacy of style unlearning for mitigating artistic imitation, we follow [32, 33] to use 20 prompts each for five famous artists and five modern artists. Evaluation uses LPIPS scores [75], where a higher  $LPIPS_e$  (on erased artists) indicates stronger removal of the target style, and a lower  $LPIPS_p$  (on preserved artists) reflects better retention of unrelated artists. We additionally use GPT-4o [76] to classify artistic styles of the generated images.  $Acc_e$  shows how often the unlearned style is still predicted – lower is better.  $Acc_p$  measures accuracy on non-erased styles – higher is better. More experimental details are provided in the Appendix. Quantitative results in Tab. 3 show that ICE

consistently achieves strong erasure efficacy with minimal impact on unintended styles compared to baselines.

Qualitative comparisons confirm this superior performance. In Fig. 3 and Fig. 6, most baselines fail to resist generating the ‘Van Gogh’ composition. ICE, in contrast, generates completely new coherent images in a different style, demonstrating a higher-fidelity erasure. We also assess the impact of erasing up to a 1000 styles, while preserving all other styles in App. Fig. 13 and find that ICE can erase up to 100 artists without affecting CLIP scores. Furthermore, Fig. 5 shows ICE’s impressive specificity in retaining untargted styles given its overlap preservation technique.

## 4.3. Object Erasure

Next, following [56], we evaluate object removal by iteratively selecting one class from 10 distinct ImageNet classes [77] as  $e$  and assessing preservation on the remaining 9. For evaluation on T2V models, we conduct per-frame classification on CogVideoX-2B generated videos and compute *Erasure Success Rate*  $ESR-k = 1 - \text{top-}k$  on the erased class and *Preservation Success Rate*  $PSR-k$  as average top- $k$  accuracy over remaining classes. As shown in Tab. 4, ICE attains the highest  $ESR-1/ESR-5$  while maintaining strong  $PSR-k$ , indicating strong suppression of the target object with minimal impact on non-target content. We further confirm strong performance on T2I models in App. Tab. 10 and App. Fig. 9. This robustness is also high-



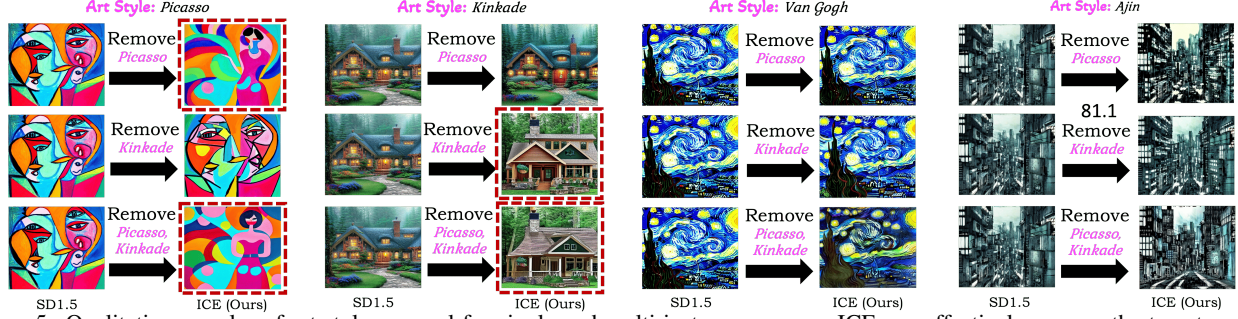


Figure 5. Qualitative samples of art style removal for single and multi-instance erasure. ICE can effectively remove the target concept (images with red borders) while preserving non-target styles with high precision and fidelity.

| Method            | Remove “Van Gogh”    |                      |                    |                    | Remove “Kelly McKernan” |                      |                    |                    |
|-------------------|----------------------|----------------------|--------------------|--------------------|-------------------------|----------------------|--------------------|--------------------|
|                   | LPIPS <sub>e</sub> ↑ | LPIPS <sub>p</sub> ↓ | Acc <sub>e</sub> ↓ | Acc <sub>p</sub> ↑ | LPIPS <sub>e</sub> ↑    | LPIPS <sub>p</sub> ↓ | Acc <sub>e</sub> ↓ | Acc <sub>p</sub> ↑ |
| SD-v1.4           | -                    | -                    | 0.95               | 0.95               | -                       | -                    | 0.80               | 0.83               |
| SLD-Medium [12]   | 0.31                 | 0.55                 | 0.95               | 0.91               | 0.39                    | 0.47                 | 0.50               | 0.79               |
| SAFREE [22]       | 0.42                 | 0.31                 | 0.35               | 0.85               | <u>0.40</u>             | 0.39                 | 0.40               | 0.78               |
| AdaVD [23]        | 0.43                 | <u>0.07</u>          | <u>0.28</u>        | 0.94               | 0.37                    | 0.06                 | <u>0.35</u>        | <b>0.97</b>        |
| CA [27]           | 0.30                 | 0.13                 | 0.65               | 0.90               | 0.22                    | 0.17                 | 0.50               | 0.76               |
| ESD [26]          | 0.40                 | 0.26                 | 0.57               | 0.89               | 0.37                    | 0.21                 | 0.81               | 0.69               |
| AdvUnlearn [48]   | 0.33                 | 0.12                 | 0.30               | 0.87               | <u>0.40</u>             | 0.13                 | 0.39               | 0.71               |
| RECE [32]         | 0.31                 | 0.08                 | 0.80               | 0.93               | 0.29                    | <u>0.04</u>          | 0.55               | 0.76               |
| UCE [33]          | 0.25                 | <b>0.05</b>          | 0.95               | <b>0.98</b>        | 0.25                    | <b>0.03</b>          | 0.80               | 0.81               |
| CURE              | <u>0.44</u>          | 0.08                 | 0.30               | 0.94               | <b>0.41</b>             | 0.09                 | <u>0.35</u>        | 0.94               |
| <b>ICE (Ours)</b> | <b>0.46</b>          | <u>0.07</u>          | <b>0.27</b>        | <u>0.96</u>        | 0.39                    | <b>0.03</b>          | <b>0.32</b>        | <u>0.95</u>        |

Table 3. Quantitative comparison of methods on the artist concept erasure task.

lighted in Fig. 6, where ICE uniquely resists an adversarial jailbreak prompt for the erased ‘car’ concept, unlike all baselines. Additional samples in App. Fig. 14 further show that ICE resists the generation of images of the synonymous form of a removed concept, demonstrating high erasure efficacy. Together, these results show ICE provides robust, specific, and generalizable object erasure for both T2I/V.

#### 4.4. Identity Erasure

We next evaluate ICE on the challenging task of identity erasure, where high inter-class similarity (e.g., between human faces) increases the risk of collateral damage. This task is critical for removing copyrighted characters or specific public figures. We select four public and stylized identities (Barack Obama, Queen Elizabeth, Mickey Mouse, and Spongebob), setting one as the target concept while preserving the others. We measure performance using average ID-Similarity via ArcFace [78] embeddings, reporting Erase Accuracy (Acc<sub>e</sub>) for the target and Preserve Accuracy (Acc<sub>p</sub>) for the non-target set. As demonstrated in Table 5, ICE shows remarkable erasure performance, successfully applying to different T2I and T2V models, as well as community SD versions [68–70], while simultaneously maintaining high accuracy for the preserved identities. This result quantitatively validates that our explicit overlap-aware framework successfully navigates this high-similarity domain, preventing collateral damage. Qualitative results in Fig. 7 demonstrate that ICE can remove the knowledge of ‘Mickey Mouse’ without affecting the concept ‘Spongebob’

| Method             | ESR-1↑       | ESR-5↑       | PSR-1↑       | PSR-5↑       |
|--------------------|--------------|--------------|--------------|--------------|
| CogVideoX-2B       | 21.62        | 5.09         | 78.38        | 94.91        |
| NegPrompt          | 48.59        | 19.79        | <u>65.37</u> | <u>88.62</u> |
| SAFREE [22]        | 61.65        | 36.41        | 53.46        | 79.17        |
| T2VUnlearning [56] | <u>92.38</u> | <u>77.09</u> | 54.03        | 82.14        |
| <b>ICE (Ours)</b>  | <b>93.10</b> | <b>82.55</b> | <b>75.41</b> | <b>89.63</b> |

Table 4. Results of ImageNet object erasure on CogVideoX-2B.

| Model           | Metric             | Obama | Elizabeth | Mickey Mouse | Spongebob | AVG  |
|-----------------|--------------------|-------|-----------|--------------|-----------|------|
| SD-v1.5         | Acc <sub>e</sub> ↓ | 0.04  | 0.09      | 0.09         | 0.12      | 0.06 |
|                 | Acc <sub>p</sub> ↑ | 0.91  | 0.87      | 0.92         | 0.83      | 0.88 |
| RealisticVision | Acc <sub>e</sub> ↓ | 0.14  | 0.12      | 0.17         | 0.17      | 0.15 |
|                 | Acc <sub>p</sub> ↑ | 0.71  | 0.78      | 0.74         | 0.69      | 0.74 |
| Chilloutmix     | Acc <sub>e</sub> ↓ | 0.21  | 0.11      | 0.07         | 0.14      | 0.13 |
|                 | Acc <sub>p</sub> ↑ | 0.80  | 0.68      | 0.72         | 0.71      | 0.73 |
| DreamShaper     | Acc <sub>e</sub> ↓ | 0.11  | 0.16      | 0.13         | 0.07      | 0.12 |
|                 | Acc <sub>p</sub> ↑ | 0.79  | 0.76      | 0.81         | 0.85      | 0.80 |
| SD-v2.1         | Acc <sub>e</sub> ↓ | 0.08  | 0.13      | 0.11         | 0.16      | 0.12 |
|                 | Acc <sub>p</sub> ↑ | 0.80  | 0.93      | 0.91         | 0.89      | 0.88 |
| CogX-2B         | Acc <sub>e</sub> ↓ | 0.06  | 0.10      | 0.09         | 0.15      | 0.10 |
|                 | Acc <sub>p</sub> ↑ | 0.85  | 0.74      | 0.83         | 0.90      | 0.83 |
| CogX-5B         | Acc <sub>e</sub> ↓ | 0.11  | 0.12      | 0.08         | 0.09      | 0.11 |
|                 | Acc <sub>p</sub> ↑ | 0.82  | 0.79      | 0.86         | 0.95      | 0.87 |

Table 5. Results of face erasure. The Acc<sub>e</sub> row reports the ID-Similarity of the target face, while the Acc<sub>p</sub> row shows the average ID-Similarity of the non-target faces.

(and vice-versa), even composing erasures to remove both simultaneously. Additional visualizations in App. Fig. 11 confirm high erasure precision, showing that untargeted concepts remain unaffected, even in cases where they share a common first name with the target identity.

#### 4.5. Unlearning Efficiency

We evaluate unlearning overheads in Tab. 6 for SD-v1.4. Prior techniques include costly training [26, 27, 48], run-time filters [12, 22, 23] do not provide preemptive weight modification, and training-free edits [32–34]. As shown in Tab. 6, ICE offers delivers strong overall performance, achieving low modification time alongside fast inference.



Figure 6. Evaluation against Ring-A-Bell prompts where, unlike baselines, ICE robustly eliminates Van Gogh’s style and the object ‘car’.

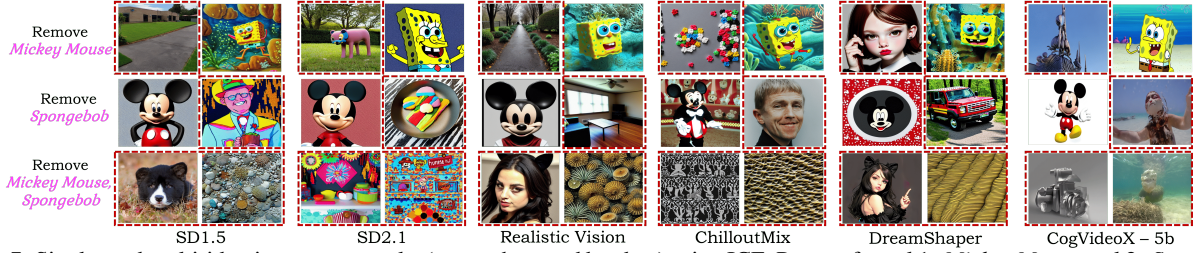


Figure 7. Single- and multi-identity erasure results (targets have red borders) using ICE. Prompt for col 1: *Mickey Mouse*; col 2: *Spongebob*.

| Method            | Mod. Time (s) | Inference Time (s/sample) | Model Mod. (%) |
|-------------------|---------------|---------------------------|----------------|
| SLD-Max [12]      | 0             | 10.34                     | 0              |
| AdaVD [23]        | 1             | 11.29                     | 0              |
| SAFREE [22]       | 0             | 10.56                     | 0              |
| ESD [26]          | ~ 4500        | 7.08                      | 94.65          |
| CA [27]           | ~ 484         | 6.31                      | 2.23           |
| AdvUnlearn [48]   | ~ 78000       | 7.07                      | 11.54          |
| UCE [33]          | ~ 1           | 7.08                      | 2.23           |
| RECE [32]         | ~ 3           | 7.12                      | 2.23           |
| CURE              | ~ 2           | 7.06                      | 2.23           |
| <b>ICE (Ours)</b> | ~ 2           | 7.07                      | 2.23           |

Table 6. Erasure efficiency comparison when removing the ‘n-dity’ concept. Evaluated on an A40 GPU for 100 iterations.

| Method                             | Acc <sub>e</sub> ↓ | Acc <sub>p</sub> ↑ |
|------------------------------------|--------------------|--------------------|
| SD-v1.4                            | 78.2               | 79.0               |
| <b>ICE (Ours)</b>                  | 1.3                | 80.3               |
| w/o Scaling                        | 0.6                | 61.4               |
| w/o Overlap Operator               | 1.3                | 69.5               |
| w/ $\mathcal{P}_e * \mathcal{P}_p$ | 1.5                | 75.1               |

Table 7. Ablating the components of ICE. We evaluate the erase accuracy (Acc<sub>e</sub>) and preserve accuracy (Acc<sub>p</sub>) on ImageNet.

## 5. Ablating ICE

We conduct an ablation study using object erasure experiments in Tab. 7. Our full method, ICE, achieves strong erasure while uniquely improving upon the original model’s utility, demonstrating surgically precise and effective unlearning. Removing anisotropic spectral scaling (w/o Scaling) results in a catastrophic drop in general performance, confirming that energy-aware basis scaling is essential for preservation. Similarly, removing the overlap-aware term (w/o Overlap Operator) leads to collateral damage, proving that modeling the semantic intersection is necessary. Finally, replacing our overlap operator with a naive matrix product ( $\mathcal{P}_e * \mathcal{P}_p$ ) partially mitigates the issue but still leaves noticeable leakage into retained content. These results show that both the importance scaling and the explicit overlap op-

erator are critical to ICE’s robust unlearning performance.

## 6. Limitations and Potential Impact

First, ICE’s efficacy is limited by the semantic granularity of the pre-trained text encoder embeddings it operates on, an issue shared by most baselines. Second, ICE’s weight modification requires white-box model access, limiting its direct use on closed-source APIs. In such cases, however, the ICE operator can still be applied in the prompt space prior to model inference. Despite these limitations, ICE advances unlearning research by establishing a mathematically grounded, efficient, and training-free paradigm for safe T2I/T2V modeling. Its transparent, modality-agnostic design enables practical applications like IP protection and model customization. Critically, its closed-form formulation promotes the interpretability and reliability essential for integration into safety-critical generative systems.

## 7. Conclusion

In this work, we introduced Instant Concept Erasure (ICE), a principled, training-free framework for one-shot concept unlearning in both T2I and T2V models. We differ from literature by mathematically formalizing the problem of semantic overlap. Next, by characterizing the intersection of erase and preserve subspaces with a closed-form analytical projector, our convex spectral objective derives a unique, optimal unlearning solution. This solution translates to a permanent, zero-overhead weight modification that precisely removes targeted artistic styles, objects, identities, and explicit content. Our extensive evaluations demonstrate that ICE achieves state-of-the-art erasure fidelity and adversarial robustness while simultaneously preserving general model quality, making the erasure of any target concept a simple case of “Now you see it, now you don’t.”

## References

- [1] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” *arXiv preprint arXiv:2010.02502*, 2020. [1](#)
- [2] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, “Glide: Towards photorealistic image generation and editing with text-guided diffusion models,” *arXiv preprint arXiv:2112.10741*, 2021. [15](#)
- [3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022. [4](#)
- [4] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, *et al.*, “Photorealistic text-to-image diffusion models with deep language understanding,” *Advances in neural information processing systems*, vol. 35, pp. 36479–36494, 2022.
- [5] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with clip latents,” *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022. [1](#)
- [6] Y. Wang, X. Chen, X. Ma, S. Zhou, Z. Huang, Y. Wang, C. Yang, Y. He, J. Yu, P. Yang, *et al.*, “Lavie: High-quality video generation with cascaded latent diffusion models,” *International Journal of Computer Vision*, vol. 133, no. 5, pp. 3059–3078, 2025. [1](#)
- [7] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng, *et al.*, “Cogvideox: Text-to-video diffusion models with an expert transformer,” *arXiv preprint arXiv:2408.06072*, 2024. [5](#)
- [8] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, *et al.*, “Imagen video: High definition video generation with diffusion models,” *arXiv preprint arXiv:2210.02303*, 2022. [1](#)
- [9] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, *et al.*, “Laion-5b: An open large-scale dataset for training next generation image-text models,” *Advances in neural information processing systems*, vol. 35, pp. 25278–25294, 2022. [1](#)
- [10] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, “The secret sharer: Evaluating and testing unintended memorization in neural networks,” in *28th USENIX security symposium (USENIX security 19)*, pp. 267–284, 2019. [1](#)
- [11] T. Lorenz, “Ai-generated porn and deepfakes raise fears of digital abuse.” <https://www.washingtonpost.com/technology/2023/02/13/ai-porn-deepfakes-women-consent/>, Feb. 2023. The Washington Post, accessed April 16, 2025. [1](#)
- [12] P. Schramowski, M. Brack, B. Deiseroth, and K. Kersting, “Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22522–22531, 2023. [1, 2, 5, 7, 8](#)
- [13] S. Andersen, K. McKernan, and K. Ortiz, “Andersen et al. v. stability ai ltd. et al., case no. 3:2023cv00201.” <https://www.courtlistener.com/docket/66732129/andersen-v-stability-ai-ltd/>, 2023. U.S. District Court for the Northern District of California, filed January 2023. [1](#)
- [14] H. H. Jiang, L. Brown, J. Cheng, M. Khan, A. Gupta, D. Workman, A. Hanna, J. Flowers, and T. 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.
- [15] K. Roose, “An a.i.-generated picture won an art prize. artists aren’t happy.” <https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html>, Sept. 2022. The New York Times, accessed April.
- [16] S. Handler, “Ai art generators hit with copyright suit over artists’ images.” <https://news.bloomberglaw.com/ip-law/ai-art-generators-hit-with-copyright-suit-over-artists-images>, Jan. 2023. Bloomberg Law, accessed April 16, 2025. [1, 16](#)
- [17] Y. Mirsky and W. Lee, “The creation and detection of deepfakes: A survey,” *ACM computing surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, 2021. [1](#)
- [18] L. Verdoliva, “Media forensics and deepfakes: an overview,” *IEEE journal of selected topics in signal processing*, vol. 14, no. 5, pp. 910–932, 2020. [1](#)
- [19] Z. Liu, G. Dou, Z. Tan, Y. Tian, and M. Jiang, “Machine unlearning in generative ai: A survey,” *arXiv preprint arXiv:2407.20516*, 2024. [1](#)
- [20] S. Kodge, G. Saha, and K. Roy, “Deep unlearning: Fast and efficient gradient-free approach to class forgetting,” *arXiv preprint arXiv:2312.00761*, 2023.
- [21] A. Huang, Z. Cai, and Z. Xiong, “A survey of machine unlearning in generative ai models: Methods, applications, security, and challenges,” *IEEE Internet of Things Journal*, 2025. [1](#)
- [22] J. Yoon, S. Yu, V. Patil, H. Yao, and M. Bansal, “Safree: Training-free and adaptive guard for safe text-to-image and video generation,” *arXiv preprint arXiv:2410.12761*, 2024. [1, 2, 5, 7, 8, 16, 17](#)
- [23] Y. Wang, O. Li, T. Mu, Y. Hao, K. Liu, X. Wang, and X. He, “Precise, fast, and low-cost concept erasure in value space: Orthogonal complement matters,” in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28759–28768, IEEE, 2025. [1, 2, 4, 5, 7, 8](#)
- [24] Y. Zhang, J. Jia, X. Chen, A. Chen, Y. Zhang, J. Liu, K. Ding, and S. Liu, “To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now,” in *European Conference on Computer Vision*, pp. 385–403, Springer, 2024. [1, 2, 5, 15](#)
- [25] M. Pham, K. O. Marshall, N. Cohen, G. Mittal, and C. Hegde, “Circumventing concept erasure methods for text-to-image generative models,” *arXiv preprint arXiv:2308.01508*, 2023. [1, 2, 15](#)



- [26] R. Gandikota, J. Materzynska, J. Fiotto-Kaufman, and D. Bau, “Erasing concepts from diffusion models,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2426–2436, 2023. 1, 2, 5, 7, 8, 15, 16, 19
- [27] N. Kumari, B. Zhang, S.-Y. Wang, E. Shechtman, R. Zhang, and J.-Y. Zhu, “Ablating concepts in text-to-image diffusion models,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22691–22702, 2023. 2, 5, 7, 8
- [28] S. Lu, Z. Wang, L. Li, Y. Liu, and A. W.-K. Kong, “Mace: Mass concept erasure in diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6430–6440, 2024. 2, 5
- [29] B. H. Lee, S. Lim, S. Lee, D. U. Kang, and S. Y. Chun, “Concept pinpoint eraser for text-to-image diffusion models via residual attention gate,” *arXiv preprint arXiv:2506.22806*, 2025. 2, 5
- [30] C. Fan, J. Liu, Y. Zhang, E. Wong, D. Wei, and S. Liu, “Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation,” *arXiv preprint arXiv:2310.12508*, 2023. 2
- [31] A. Heng and H. Soh, “Selective amnesia: A continual learning approach to forgetting in deep generative models,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 17170–17194, 2023. 1, 2, 5
- [32] C. Gong, K. Chen, Z. Wei, J. Chen, and Y.-G. Jiang, “Reliable and efficient concept erasure of text-to-image diffusion models,” in *European Conference on Computer Vision*, pp. 73–88, Springer, 2024. 1, 2, 4, 5, 6, 7, 8, 14, 16, 19
- [33] R. Gandikota, H. Orgad, Y. Belinkov, J. Materzyńska, and D. Bau, “Unified concept editing in diffusion models,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5111–5120, 2024. 4, 5, 6, 7, 8, 14, 15, 19
- [34] S. D. Biswas, A. Roy, and K. Roy, “Cure: Concept unlearning via orthogonal representation editing in diffusion models,” *arXiv preprint arXiv:2505.12677*, 2025. 1, 2, 5, 7, 16
- [35] J. Ho and T. Salimans, “Classifier-free diffusion guidance,” *arXiv preprint arXiv:2207.12598*, 2022. 1, 3, 15
- [36] G. Saha, I. Garg, and K. Roy, “Gradient projection memory for continual learning,” *arXiv preprint arXiv:2103.09762*, 2021. 1, 3
- [37] S. Kodge, D. Ravikumar, G. Saha, and K. Roy, “Sap: Corrective machine unlearning with scaled activation projection for label noise robustness,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 17930–17937, 2025. 1
- [38] G. Saha and K. Roy, “Continual learning with scaled gradient projection,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, pp. 9677–9685, 2023. 2, 3
- [39] J. Rando, D. Paleka, D. Lindner, L. Heim, and F. Tramèr, “Red-teaming the stable diffusion safety filter,” *arXiv preprint arXiv:2210.04610*, 2022. 2
- [40] Y. Cai, S. Yin, Y. Wei, C. Xu, W. Mao, F. Juefei-Xu, S. Chen, and Y. Wang, “Ethical-lens: Curbing malicious usages of open-source text-to-image models,” *Patterns*, vol. 6, no. 3, 2025. 2
- [41] M. Lyu, Y. Yang, H. Hong, H. Chen, X. Jin, Y. He, H. Xue, J. Han, and G. Ding, “One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7559–7568, 2024. 2
- [42] M. Pham, K. O. Marshall, C. Hegde, and N. Cohen, “Robust concept erasure using task vectors,” *arXiv preprint arXiv:2404.03631*, 2024.
- [43] C.-P. Huang, K.-P. Chang, C.-T. Tsai, Y.-H. Lai, F.-E. Yang, and Y.-C. F. Wang, “Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers,” in *European Conference on Computer Vision*, pp. 360–376, Springer, 2024. 2
- [44] X. Li, Y. Yang, J. Deng, C. Yan, Y. Chen, X. Ji, and W. Xu, “Safegen: Mitigating unsafe content generation in text-to-image models,” *arXiv e-prints*, pp. arXiv–2404, 2024. 2
- [45] G. Zhang, K. Wang, X. Xu, Z. Wang, and H. Shi, “Forget-me-not: Learning to forget in text-to-image diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1755–1764, 2024.
- [46] Y. Wu, S. Zhou, M. Yang, L. Wang, H. Chang, W. Zhu, X. Hu, X. Zhou, and X. Yang, “Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 8496–8504, 2025.
- [47] H. Li, C. Shen, P. Torr, V. Tresp, and J. Gu, “Self-discovering interpretable diffusion latent directions for responsible text-to-image generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12006–12016, 2024. 5
- [48] Y. Zhang, X. Chen, J. Jia, Y. Zhang, C. Fan, J. Liu, M. Hong, K. Ding, and S. Liu, “Defensive unlearning with adversarial training for robust concept erasure in diffusion models,” *Advances in neural information processing systems*, vol. 37, pp. 36748–36776, 2024. 2, 7, 8
- [49] A. Das, V. Duddu, R. Zhang, and N. Asokan, “Espresso: Robust concept filtering in text-to-image models,” *arXiv preprint arXiv:2404.19227*, 2024. 2
- [50] Y.-H. Park, S. Yun, J.-H. Kim, J. Kim, G. Jang, Y. Jeong, J. Jo, and G. Lee, “Direct unlearning optimization for robust and safe text-to-image models,” *arXiv preprint arXiv:2407.21035*, 2024. 2
- [51] H. Zhang, Y. He, and H. Chen, “Steerdiff: Steering towards safe text-to-image diffusion models,” *arXiv preprint arXiv:2410.02710*, 2024. 2
- [52] H. Orgad, B. Kawar, and Y. Belinkov, “Editing implicit assumptions in text-to-image diffusion models,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7053–7061, 2023. 2
- [53] Z.-Y. Chin, C.-M. Jiang, C.-C. Huang, P.-Y. Chen, and W.-C. Chiu, “Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts,” *arXiv preprint arXiv:2309.06135*, 2023. 2, 5, 15
- [54] Y. Yang, R. Gao, X. Wang, T.-Y. Ho, N. Xu, and Q. Xu, “Mma-diffusion: Multimodal attack on diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7737–7746, 2024. 2, 5, 15

- [55] S. Liu and Y. Tan, “Unlearning concepts from text-to-video diffusion models,” *arXiv preprint arXiv:2407.14209*, 2024. [2](#)
- [56] X. Ye, S. Cheng, Y. Wang, Y. Xiong, and Y. Li, “T2vunlearning: A concept erasing method for text-to-video diffusion models,” *arXiv preprint arXiv:2505.17550*, 2025. [2](#), [6](#), [7](#), [16](#), [17](#)
- [57] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*, pp. 8748–8763, PmLR, 2021. [3](#)
- [58] J. Ni, G. H. Abrego, N. Constant, J. Ma, K. B. Hall, D. Cer, and Y. Yang, “Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models,” *arXiv preprint arXiv:2108.08877*, 2021.
- [59] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020. [3](#)
- [60] A. Ben-Israel, “Projectors on intersection of subspaces,” *Contemporary Mathematics*, vol. 636, pp. 41–50, 2015. [3](#), [12](#)
- [61] R. Piziak, P. L. Odell, and R. Hahn, “Constructing projections on sums and intersections,” *Computers & Mathematics with Applications*, vol. 37, no. 1, pp. 67–74, 1999. [3](#), [12](#), [13](#)
- [62] K. Meng, A. S. Sharma, A. Andonian, Y. Belinkov, and D. Bau, “Mass-editing memory in a transformer,” *arXiv preprint arXiv:2210.07229*, 2022. [4](#)
- [63] Y.-L. Tsai, C.-Y. Hsu, C. Xie, C.-H. Lin, J.-Y. Chen, B. Li, P.-Y. Chen, C.-M. Yu, and C.-Y. Huang, “Ring-a-bell! how reliable are concept removal methods for diffusion models?,” *arXiv preprint arXiv:2310.10012*, 2023. [5](#), [15](#)
- [64] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” *Advances in neural information processing systems*, vol. 30, 2017. [5](#)
- [65] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, “Clipscore: A reference-free evaluation metric for image captioning,” *arXiv preprint arXiv:2104.08718*, 2021. [5](#)
- [66] CompVis, S. AI, and LAION, “Stable diffusion v1-5 model card,” <https://huggingface.co/runwayml/stable-diffusion-v1-5>, 2022. Accessed: 2025-09-19. [4](#)
- [67] Stability AI, “Stable diffusion 2.1.” Hugging Face model card, 2022. Accessed 2025-11-03. [4](#)
- [68] SG161222, “Realistic vision v5.1 (novae).” [https://huggingface.co/SG161222/Realistic\\_Vision\\_V5.1\\_noVAE](https://huggingface.co/SG161222/Realistic_Vision_V5.1_noVAE), 2023. [4](#), [7](#)
- [69] Lykon, “Dreamshaper.” <https://huggingface.co/Lykon/DreamShaper>, 2023. [4](#)
- [70] swl-models, “Chilloutmix.” <https://huggingface.co/swl-models/chilloutmix>, 2023. [4](#), [7](#)
- [71] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part v 13*, pp. 740–755, Springer, 2014. [5](#)
- [72] P. Bedapudi, “Nudenet: Neural nets for nudity classification, detection and selective censoring,” 2019. [5](#)
- [73] J. Dai, T. Chen, X. Wang, Z. Yang, T. Chen, J. Ji, and Y. Yang, “Safesora: Towards safety alignment of text2video generation via a human preference dataset,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 17161–17214, 2024. [5](#), [16](#)
- [74] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit, *et al.*, “Vbench: Comprehensive benchmark suite for video generative models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024. [6](#)
- [75] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018. [6](#)
- [76] OpenAI, “Hello GPT-4o,” May 2024. Accessed: 2025-11-10. [6](#), [16](#)
- [77] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009. [6](#)
- [78] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019. [7](#)
- [79] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012. [14](#)
- [80] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, “Null-text inversion for editing real images using guided diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6038–6047, 2023. [15](#)
- [81] G. Li, K. Chen, S. Zhang, J. Zhang, and T. Zhang, “Art: Automatic red-teaming for text-to-image models to protect benign users,” *arXiv preprint arXiv:2405.19360*, 2024. [15](#)
- [82] K. Chen, Z. Wei, J. Chen, Z. Wu, and Y.-G. Jiang, “Gcma: Generative cross-modal transferable adversarial attacks from images to videos,” in *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 698–708, 2023. [15](#)
- [83] M. Kim, H. Lee, B. Gong, H. Zhang, and S. J. Hwang, “Automatic jailbreaking of the text-to-image generative ai systems,” *arXiv preprint arXiv:2405.16567*, 2024. [15](#)
- [84] Y. Miao, Y. Zhu, L. Yu, J. Zhu, X.-S. Gao, and Y. Dong, “T2vsafetybench: Evaluating the safety of text-to-video generative models,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 63858–63872, 2024. [16](#)

## Appendix

In this appendix, we provide detailed proofs, additional experimental results, and implementation details to supplement the main paper. We begin with the formal mathematical derivations for our method: Sec. A provides the proof for the closed-form overlap projector, Sec. B proves the L-smoothness of our spectral objective, Sec. C proves its strict convexity, and Sec. D details the full derivation of the closed-form ICE solution. Following this, Sec. E discusses additional implementation details, while Sec. F provides extended experimental results, including comprehensive object erasure benchmarks, sensitive-concept erasure results and additional qualitative visualizations. Finally, Sec. G provides a complete list of licenses for all models and datasets used in this work.

### A. Overlap Operator Proof

**Proposition 3.** *The unique overlap projector,  $\mathcal{P}_{e \cap p}$ , onto the intersection of subspaces  $\mathcal{S}_e$  and  $\mathcal{S}_p$  is given in closed-form by:*

$$\mathcal{P}_{e \cap p} = 2\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p \quad (11)$$

where  $\mathcal{P}_e$  and  $\mathcal{P}_p$  are the projectors onto  $\mathcal{S}_e$  and  $\mathcal{S}_p$  respectively, and  $^\dagger$  denotes the Moore-Penrose pseudo inverse.

**Preliminaries.** To facilitate the subsequent proof, we first review the fundamental properties of projection operators, specifically their construction for the linear sum and intersection of subspaces, as established in [60]. We begin by defining the requisite notation:

If  $M$  is a subspace of  $\mathbb{S}^n$ , we write  $\mathcal{P}_M$  for the unique projection onto  $M$ . We denote  $M = \text{Span}(\mathcal{P}_M)$  and  $M^\perp = \text{Span}^\perp(\mathcal{P}_M)$ . We further find from [61] that if say  $A$  and  $B$  are projections on  $\mathbb{S}^n$ ,

$$\text{Span}(A + B) = \text{Span}(A) + \text{Span}(B), \quad (12)$$

By Eq. (12),

$$\text{Span}(A) \subseteq \text{Span}(A) + \text{Span}(B) = \text{Span}(A + B) \quad (13)$$

With this property for subspaces established, we now proceed to the formal proof.

*Proof.* We wish to find the span of the overlap projection, i.e., to characterize  $\text{Span}(\mathcal{P}_{e \cap p}) = \mathcal{S}_e \cap \mathcal{S}_p$  directly from  $\mathcal{P}_e$  and  $\mathcal{P}_p$ .

The proof follows the style of [61], adapting the notation. We begin by proving commutativity property:

$$2\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p = 2\mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e \quad (14)$$

To do this, we show that their difference is zero. We start by adding and subtracting the term  $\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e$ :

$$\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p - \mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e \quad (15)$$

$$= \mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p + \mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e -$$

$$[\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e + \mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e] \quad (16)$$

$$= \mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger (\mathcal{P}_e + \mathcal{P}_p) - (\mathcal{P}_e + \mathcal{P}_p)(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e \quad (17)$$

From Eq. 13, we have  $\text{Span}(\mathcal{P}_e) \subseteq \text{Span}(\mathcal{P}_e + \mathcal{P}_p)$ . In other words,  $\mathcal{S}_e \subseteq \mathcal{S}_e + \mathcal{S}_p$ . For any two projectors with nested subspaces  $\mathcal{S}_1 \subseteq \mathcal{S}_2$ , the property  $\mathcal{P}_{\mathcal{S}_2} \mathcal{P}_{\mathcal{S}_1} = \mathcal{P}_{\mathcal{S}_1}$  holds. This is because  $\mathcal{P}_{\mathcal{S}_1}$  projects any vector into  $\mathcal{S}_1$ , which is already in  $\mathcal{S}_2$ , so the subsequent projection  $\mathcal{P}_{\mathcal{S}_2}$  has no effect. Therefore,

$$\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger (\mathcal{P}_e + \mathcal{P}_p) = \mathcal{P}_e \quad (18)$$

$$(\mathcal{P}_e + \mathcal{P}_p)(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e = \mathcal{P}_e \quad (19)$$

Substituting these results back into our expression for the difference (Eq. 17) yields  $\mathcal{P}_e - \mathcal{P}_e = 0$ . Thus, the difference for RHS in Eq. 17 is zero, and we have proven the commutativity, i.e

$$\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p - \mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e = 0 \quad (20)$$

$$\implies \mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p = \mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e \quad (21)$$

$$\implies 2\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p = 2\mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e \quad (22)$$

Next, we argue that both sides of Eq. (22) equal  $\mathcal{P}_{e \cap p}$ . Let

$$H = \mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p + \mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e \quad (23)$$

$$= 2\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p \quad (24)$$

$$= 2\mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e \quad (25)$$

$$(26)$$

We first show that  $\text{Span}(H) \subseteq \mathcal{S}_e \cap \mathcal{S}_p$ .

$$H\mathcal{P}_p = [2\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p]\mathcal{P}_p \quad (27)$$

$$= 2\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p^2 \quad (28)$$

$$= 2\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p \text{ (by projector property [61])} \quad (29)$$

$$= H \quad (30)$$

The property  $H\mathcal{P}_p = H$  implies that projecting into  $H\mathcal{P}_p$  results in projection onto  $H$  itself. This means that  $\text{Span}(H)$  refers to a nested subspace within  $\text{Span}(\mathcal{P}_p)$ . In other words,  $\text{Span}(H) \subseteq \text{Span}(\mathcal{P}_p) = \mathcal{S}_p$ .

Similarly, we prove  $H\mathcal{P}_e = H$  which implies  $\text{Span}(H) \subseteq \text{Span}(\mathcal{P}_e) = \mathcal{S}_e$ . Since  $\text{Span}(H)$  is a subspace of both  $\mathcal{S}_e$  and  $\mathcal{S}_p$ , it must be a subspace of their intersection:  $\text{Span}(H) \subseteq \mathcal{S}_e \cap \mathcal{S}_p$ .



Because  $\text{Span}(H) \subseteq \mathcal{S}_e \cap \mathcal{S}_p$  and  $\mathcal{P}_{e \cap p}$  is the projector onto  $\mathcal{S}_e \cap \mathcal{S}_p$ , it follows that  $H\mathcal{P}_{e \cap p} = H$ . Hence, we write:

$$H = H\mathcal{P}_{e \cap p} \quad (31)$$

$$= [\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p + \mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e] \mathcal{P}_{e \cap p} \quad (32)$$

$$= \mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p \mathcal{P}_{e \cap p} + \mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e \mathcal{P}_{e \cap p} \quad (33)$$

$$= \mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_{e \cap p} + \mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_{e \cap p} \quad (34)$$

$$= [\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger + \mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger] \mathcal{P}_{e \cap p} \quad (35)$$

$$= [(\mathcal{P}_e + \mathcal{P}_p)(\mathcal{P}_e + \mathcal{P}_p)^\dagger] \mathcal{P}_{e \cap p} \quad (36)$$

$$= \mathcal{P}_{e \cap p} \quad (37)$$

This last equality follows because  $\text{Span}(\mathcal{P}_{e \cap p}) \subseteq \text{Span}(\mathcal{P}_e + \mathcal{P}_p)$ .

Thus from Eq. 37,  $\mathcal{P}_{e \cap p} = H = 2\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p$ . The preceding proof, based on the properties of projectors from [61], establishes the span and mathematical form of the subspace intersection  $\mathcal{P}_{e \cap p}$ .

In our practical implementation, we create  $\mathcal{P}_{e \cap p}$  with our energy-scaled operators, using anisotropic scaling. This is because not all basis components are equally important for a given concept, and hence we take into consideration the non-uniform energy distribution of each basis when creating the unlearning operators. To reflect this in practice, the scaling function assigns maximal salience to the principal directions ( $\lambda_i = 1$ ), while others are attenuated relative to this maximum. This energy attenuation maintains all the properties we desire since the scaling function produces strictly positive importance scores ( $\lambda_i > 0$ ) for all non-zero singular values. Hence, our scaled operators preserve the original span of the concepts; only the energy along basis directions is attenuated. ■

## B. Proof of L-Smoothness for the Spectral Objective

**Proposition 4.** *The spectral objective function  $\mathcal{L}(\mathbf{x}_{ice})$  is L-smooth, i.e., its gradient is Lipschitz continuous.*

*Proof.* We must show there exists a constant  $K \geq 0$  such that for any two vectors  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ , the following inequality holds:  $\|\nabla \mathcal{L}(\mathbf{x}) - \nabla \mathcal{L}(\mathbf{y})\|_2 \leq K \cdot \|\mathbf{x} - \mathbf{y}\|_2$ .

The gradient of the objective function is:

$$\nabla \mathcal{L}(\mathbf{x}_{ice}) = 2(\mathbf{x}_{ice} - \mathbf{x}\mathcal{P}_e) + 2(\mathbf{x}_{ice}\mathcal{P}_{e \cap p})\mathcal{P}_{e \cap p}^T$$

Let's evaluate the difference of the gradients at two points,  $\mathbf{x}$  and  $\mathbf{y}$ . The constant terms involving the initial point  $\mathbf{x}$

cancel out:

$$\begin{aligned} \nabla \mathcal{L}(\mathbf{x}) - \nabla \mathcal{L}(\mathbf{y}) &= (2\mathbf{x} + 2\mathbf{x}\mathcal{P}_{e \cap p}\mathcal{P}_{e \cap p}^T) - \\ &\quad (2\mathbf{y} + 2\mathbf{y}\mathcal{P}_{e \cap p}\mathcal{P}_{e \cap p}^T) \\ &= 2(\mathbf{x} - \mathbf{y}) + 2(\mathbf{x} - \mathbf{y})\mathcal{P}_{e \cap p}\mathcal{P}_{e \cap p}^T \\ &= 2(\mathbf{x} - \mathbf{y})(\mathcal{I} + \mathcal{P}_{e \cap p}\mathcal{P}_{e \cap p}^T) \end{aligned}$$

Now, we take the L2 norm of both sides. Using the property of induced matrix norms,  $\|\mathbf{v}\mathcal{A}\|_2 \leq \|\mathbf{v}\|_2 \cdot \|\mathcal{A}\|_2$  (By Triangle Inequality), we have:

$$\|\nabla \mathcal{L}(\mathbf{x}) - \nabla \mathcal{L}(\mathbf{y})\|_2 = \|2(\mathbf{x} - \mathbf{y})(\mathcal{I} + \mathcal{P}_{e \cap p}\mathcal{P}_{e \cap p}^T)\|_2 \quad (38)$$

$$\leq 2 \cdot \|\mathcal{I} + \mathcal{P}_{e \cap p}\mathcal{P}_{e \cap p}^T\|_2 \cdot \|\mathbf{x} - \mathbf{y}\|_2 \quad (39)$$

We can define the Lipschitz constant  $K$  as:

$$K = 2 \cdot \|\mathcal{I} + \mathcal{P}_{e \cap p}\mathcal{P}_{e \cap p}^T\|_2 \quad (40)$$

Since  $\mathcal{I}$  and the projector  $\mathcal{P}_{e \cap p}$  are fixed with finite entries, their operator norms are finite, so a finite Lipschitz constant  $K$  exists. By the triangle inequality,

$$K \leq 2(\|\mathcal{I}\|_2 + \|\mathcal{P}_{e \cap p}\mathcal{P}_{e \cap p}^T\|_2). \quad (41)$$

Let  $\mathcal{P}_{e \cap p} = U\Lambda U^\top$  with  $\Lambda = \text{diag}(\lambda_i)$ . Then

$$\|\mathcal{P}_{e \cap p}\mathcal{P}_{e \cap p}^T\|_2 = \|U\Lambda^2 U^\top\|_2 \quad (42)$$

$$= \text{Tr}((U\Lambda^2 U^\top)^\top (U\Lambda^2 U^\top)) \quad (43)$$

$$= \text{Tr}(U\Lambda^4 U^\top) \quad (44)$$

$$= \text{Tr}(U^\top U \Lambda^4) \text{ (by cyclic property of trace)} \quad (45)$$

$$= \text{Tr}(\Lambda^4) \quad (46)$$

$$\leq d \text{ (Since } \max_i \lambda_i^2 \leq 1) \quad (47)$$

where  $d$  is the dimension of matrix  $\Lambda^4$ . As  $\|\mathcal{I}\|_2 = 1$  and  $\text{Tr}(\Lambda^4) \leq d$ , we obtain

$$K \leq 2(1 + d). \quad (48)$$

Thus, the gradient is Lipschitz continuous with constant  $K \leq 2(1 + d)$ . ■

## C. Proof of Convexity for the Spectral Objective

**Proposition 5.** *The spectral objective function  $\mathcal{L}(\mathbf{x}_{ice})$  as defined in Eq. (6) is strictly convex.*

*Proof.* A twice-differentiable function is strictly convex if its Hessian matrix is positive definite. We will compute the

Hessian of  $\mathcal{L}(\mathbf{x}_{ice})$  and show that it meets this criterion. Note that  $\mathbf{x}_{ice}$  is a 1-d vector.

First, we restate the objective function for clarity:

$$\mathcal{L}(\mathbf{x}_{ice}) = \|\mathbf{x}_{ice} - \mathbf{x}\mathcal{P}_e\|_2^2 + \|\mathbf{x}_{ice}\mathcal{P}_{e\cap p}\|_2^2 \quad (49)$$

Expanding the squared norms gives:

$$\mathcal{L}(\mathbf{x}_{ice}) = (\mathbf{x}_{ice} - \mathbf{x}\mathcal{P}_e)(\mathbf{x}_{ice} - \mathbf{x}\mathcal{P}_e)^T + \quad (50)$$

$$(\mathbf{x}_{ice}\mathcal{P}_{e\cap p})(\mathbf{x}_{ice}\mathcal{P}_{e\cap p})^T \quad (51)$$

The gradient of  $\mathcal{L}$  with respect to the row vector  $\mathbf{x}_{ice}$  is:

$$\nabla_{\mathbf{x}_{ice}} \mathcal{L} = 2(\mathbf{x}_{ice} - \mathbf{x}\mathcal{P}_e) + 2(\mathbf{x}_{ice}\mathcal{P}_{e\cap p})\mathcal{P}_{e\cap p}^T \quad (52)$$

The Hessian matrix,  $\mathcal{H}$ , is the derivative of the gradient with respect to  $\mathbf{x}_{ice}$ . Differentiating the gradient term-by-term, we find:

$$\begin{aligned} \frac{\partial}{\partial \mathbf{x}_{ice}} (2\mathbf{x}_{ice}) &= 2\mathcal{I} \\ \frac{\partial}{\partial \mathbf{x}_{ice}} (-2\mathbf{x}\mathcal{P}_e) &= 0 \\ \frac{\partial}{\partial \mathbf{x}_{ice}} (2\mathbf{x}_{ice}\mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T) &= 2\mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T \end{aligned}$$

Summing these terms gives the Hessian:

$$\mathcal{H} = \nabla_{\mathbf{x}_{ice}}^2 \mathcal{L} = 2\mathcal{I} + 2\mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T \quad (53)$$

To prove that  $\mathcal{L}$  is strictly convex, we must show that  $\mathcal{H}$  is positive definite, i.e.,  $\mathbf{v}^T \mathcal{H} \mathbf{v} > 0$  for any non-zero vector  $\mathbf{v} \in \mathbb{R}^d$ .

$$\begin{aligned} \mathbf{v}^T \mathcal{H} \mathbf{v} &= \mathbf{v}^T (2\mathcal{I} + 2\mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T) \mathbf{v} \\ &= 2\mathbf{v}^T \mathcal{I} \mathbf{v} + 2\mathbf{v}^T \mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T \mathbf{v} \\ &= 2(\mathbf{v}^T \mathbf{v}) + 2(\mathcal{P}_{e\cap p}^T \mathbf{v})^T (\mathcal{P}_{e\cap p}^T \mathbf{v}) \\ &= 2\|\mathbf{v}\|_2^2 + 2\|\mathcal{P}_{e\cap p}^T \mathbf{v}\|_2^2 \end{aligned}$$

The first term,  $2\|\mathbf{v}\|_2^2$ , is strictly positive for any non-zero vector  $\mathbf{v}$ . The second term,  $2\|\mathcal{P}_{e\cap p}^T \mathbf{v}\|_2^2$ , is a squared norm and is therefore non-negative ( $\geq 0$ ).

The sum of a strictly positive term and a non-negative term is strictly positive. Therefore,  $\mathbf{v}^T \mathcal{H} \mathbf{v} > 0, \forall \mathbf{v} \neq \mathbf{0}$ .

Since the Hessian matrix  $\mathcal{H}$  is positive definite, the objective function  $\mathcal{L}(\mathbf{x}_{ice})$  is strictly convex. This guarantees that any stationary point is the unique global minimum. ■

## D. Deriving the Closed-Form Solution for ICE

**Proposition 6.** *Given a concept direction vector  $\mathbf{x}$  and projection operators  $\mathcal{P}_e$ ,  $\mathcal{P}_p$ , and the overlap operator  $\mathcal{P}_{e\cap p}$  the vector  $\mathbf{x}_{ice}$  that minimizes the objective  $\mathcal{L}(\mathbf{x}_{ice})$  in Equation 6 is given by:*

$$\mathbf{x}_{ice} = \mathbf{x}(\mathcal{P}_e)(\mathcal{I} + \mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T)^{-1} \quad (54)$$

*Proof.* To find the minimum of the convex function  $\mathcal{L}(\mathbf{x}_{ice})$ , we compute its gradient with respect to the row vector  $\mathbf{x}_{ice}$  and set it to zero.

$$\mathcal{L}(\mathbf{x}_{ice}) = \|\mathbf{x}_{ice} - \mathbf{x}\mathcal{P}_e\|_2^2 + \|\mathbf{x}_{ice}\mathcal{P}_{e\cap p}\|_2^2 \quad (55)$$

Using the identity  $\nabla_{\mathcal{X}} \|\mathcal{X}\mathcal{A} - \mathcal{B}\|_2^2 = 2(\mathcal{X}\mathcal{A} - \mathcal{B})\mathcal{A}^T$ , the gradient is:

$$\nabla_{\mathbf{x}_{ice}} \mathcal{L}(\mathbf{x}_{ice}) = 2(\mathbf{x}_{ice} - \mathbf{x}\mathcal{P}_e) + 2(\mathbf{x}_{ice}\mathcal{P}_{e\cap p})(\mathcal{P}_{e\cap p})^T \quad (56)$$

Setting  $\nabla_{\mathbf{x}_{ice}} \mathcal{L} = 0$  and dividing by 2 yields:

$$(\mathbf{x}_{ice} - \mathbf{x}\mathcal{P}_e) + (\mathbf{x}_{ice}\mathcal{P}_{e\cap p})(\mathcal{P}_{e\cap p}^T) = 0 \quad (57)$$

$$\implies \mathbf{x}_{ice} - \mathbf{x}\mathcal{P}_e + \mathbf{x}_{ice}\mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T = 0 \quad (58)$$

We rearrange the equation to isolate terms involving  $\mathbf{x}_{ice}$ :

$$\mathbf{x}_{ice}(\mathcal{I} + \mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T) = \mathbf{x}(\mathcal{P}_e) \quad (59)$$

The term  $(\mathcal{I} + \mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T)$  is guaranteed to be invertible. By construction, the matrix  $\mathcal{A} = \mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T$  is positive semi-definite, as for any non-zero vector  $\mathbf{z}$ ,

$$\mathbf{z}^T \mathcal{A} \mathbf{z} = \mathbf{z}^T (\mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T) \mathbf{z} \quad (60)$$

$$= (\mathcal{P}_{e\cap p}^T \mathbf{z})^T (\mathcal{P}_{e\cap p}^T \mathbf{z}) \quad (61)$$

$$= \|\mathcal{P}_{e\cap p}^T \mathbf{z}\|_2^2 \quad (62)$$

$$\geq 0 \quad (63)$$

The sum of the identity matrix  $\mathcal{I}$  (which is positive definite) and a positive semi-definite matrix is always positive definite. A positive definite matrix has all strictly positive eigenvalues and is therefore invertible [79]. Right-multiplying by the inverse of this matrix isolates  $\mathbf{x}_{ice}$ :

$$\mathbf{x}_{ice} = \mathbf{x}(\mathcal{P}_e)(\mathcal{I} + \mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T)^{-1} \quad (64)$$

This concludes the proof. ■

## E. Additional Discussions

### E.1. Subspace construction and Prompt Templates

For each target concept, we construct the embedding basis using concise prompt templates that substitute the concept into common forms: “picture of/by [placeholder]” “photo of/by [placeholder]” “image of/by [placeholder]” “portrait of/by [placeholder]”, “painting of/by [placeholder]”. This is consistent with prior works [32, 33]. Empirically, we observe using 3-5 diverse prompts suffices to construct a stable and expressive embedding basis. For unsafe content erasure, we adopt the target prompt “violence, nudity, harm”, following established protocol in [33] for fair comparison.

## E.2. Unconditional Embedding as a Preserve Set

In the main paper, we state that the preserve set  $\mathcal{S}_p$  is designed to represent the ‘broad domain of all other possible concepts’ and that we use the unconditional embedding “ ” for this. Here, we elaborate on this choice.

In modern diffusion models, generation is typically guided using Classifier-Free Guidance (CFG) [35]. This technique requires two inputs at each denoising step: the text embedding for the desired prompt (e.g., “a photo of a cat”) and an unconditional embedding. This unconditional embedding is most commonly the embedding of an empty string (“ ”), representing the model’s prior in the absence of a specific concept [2, 80]. It encapsulates the generic, shared features learned from the entire training dataset, rather than any single specific concept [26, 33].

The goal of unlearning is to remove a concept  $e$  (e.g., “Van Gogh”) while preserving all other concepts  $p$ . However, manually defining  $p$  by listing every other concept (e.g., “Monet”, “dog”, “tree”, “painting”, etc.) may be computationally intractable and conceptually impossible for every target erase concept. The unconditional embedding provides a powerful and efficient proxy for this broad domain of all other possible concepts. By setting  $\mathcal{S}_p$  as the subspace defined by this generic “ ” embedding, we are effectively defining the preserve set as the model’s average, generic, and common-sense knowledge. When our method then calculates the intersection  $\mathcal{S}_e \cap \mathcal{S}_p$ , it identifies the features of the erase concept  $e$  that are shared with this generic, average representation. For example, when erasing “Van Gogh”, the components of its embedding that also mean “painting” or “art” (which are captured in the average, unconditional embedding) are identified as the intersection. Our objective then explicitly preserves these shared components. This ensures that ICE only ablates the unique, identifying features of the target concept while leaving the shared, general semantics (like “painting”) unharmed, which is the key to preventing the collateral damage seen in naive methods.

## E.3. ICE Through the Lens of Dissociation via Set Difference

As discussed in the main paper, the erase subspace  $\mathcal{S}_e$  (e.g., “Van Gogh”) is generally *not* orthogonal to the preserve subspace  $\mathcal{S}_p$  (e.g., “painting” or the unconditional embedding “ ”). Naïve orthogonal removal that ignores  $\mathcal{S}_e \cap \mathcal{S}_p$  will inevitably suppress shared directions and hurt untargeted content quality. The practical failure to explicitly model and protect this overlap is the source of fidelity-robustness trade-offs noted in the main text.

Let the anisotropically scaled conditioning operators be

$$\mathcal{P}_e = \mathcal{U}_e \Lambda_e \mathcal{U}_e^\top, \quad \mathcal{P}_p = \mathcal{U}_p \Lambda_p \mathcal{U}_p^\top,$$

where the diagonal weights  $\Lambda_e, \Lambda_p$  emphasize high-energy, concept-defining directions. To preserve shared semantics,

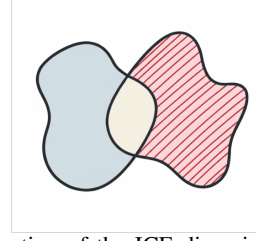


Figure 8. Visualization of the ICE dissociation operation. The hatched region denotes

ICE isolates the exact intersection via:

$$\mathcal{P}_{e \cap p} = 2 \mathcal{P}_e (\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p$$

The dissociation operator we apply is then the *overlap-aware* projector

$$\mathcal{P}_{ice} = \mathcal{P}_e - \mathcal{P}_{e \cap p}$$

which intuitively isolates the ‘ $e$ ’ minus ‘ $e \cap p$ ’ set, as visualized in Fig. 8. This leads to unlearning the highly target concept specific erase-only directions while leaving the shared intersection untouched.

The ICE Spectral Objective is a functional implementation of this principle. The Preservation Term, which is dependent on our characterization of the intersection operator, acts as a strong regularizer. It penalizes any unlearning solution that attempts to modify this shared subspace. The resulting closed-form solution is therefore an operator that effectively isolates the discriminative erase concept space from shared semantic components.

The practical effect of this targeted dissociation is a measurable reduction in the semantic similarity between the model’s representations of the erase and preserve concepts. We quantify this by computing the average cosine similarity between the embeddings of erase concepts ( $e$ ) and their semantically-related preserve concepts ( $p$ ) before and after our one-shot weight modification. As shown in Fig. 1, ICE significantly reduces the original embedding similarity. This demonstrates that our method successfully dissociates the entangled representations, which in turn prevents the collateral damage observed otherwise, and leads to the strong preservation performance reported in the main paper.

## E.4. Red-Teaming Attacks

As safety mechanisms become more prevalent, recent works have explored adversarial attacks [81, 82] and jail-breaking [83] to evaluate the robustness of unlearned T2I models. White-box attacks like [24, 25, 53] exploit the classification capacity or prompt-conditioned behavior of diffusion models to revive erased concepts. In contrast, black-box methods like [63] use evolutionary algorithms to generate adversarial prompts or exploit text embeddings and multimodal inputs to bypass safeguards [54]. These tools



reveal critical vulnerabilities in concept removal approaches when deployed in unrestricted environments and while several unlearning frameworks partially mitigate these attacks, very few are robust across all threat types. We therefore explicitly evaluate unlearning method robustness under both white-box and black-box red-teaming attacks drawn from these prior works.

### E.5. Setting the NudeNet Threshold

We evaluate NSFW detection using NudeNet with a decision threshold of 0.6. This choice follows recent practice in safety filtering [32, 34], where this value has been adopted to better capture borderline NSFW content. This threshold ensures compatibility with safety-sensitive applications by being sufficiently conservative. Importantly, for fairness and consistency, all methods in our evaluation, including baselines, have been assessed using this same threshold.

### E.6. Details on Evaluation Dataset Gen

We adopt the Gen dataset for unsafe content erasure assessment from prior work [56], to evaluate unlearning robustness against LLM-refined prompts. To create the dataset, prompt augmentation was applied to the concept of ‘nudity’ and 100 evaluation prompts were generated following the format requirements of CogVideoX. Example prompts are provided in Tab. 8.

### E.7. Details on SafeSora Evaluation

To quantitatively evaluate the unsafe concept filtering ability on T2V, we choose SafeSora [73], which contains 600 toxic textual prompts across 12 toxic concepts as our testbed. Following [22], we further select 5 representative categories within 12 concepts, and thus construct a safe video generation benchmark with 296 examples. For the evaluation metrics, we follow the automatic evaluation via ChatGPT proposed by T2VSafetybench [84]. We input sampled 16 video frames along with the same prompt design presented in T2VSafetybench to GPT-4o [76] for binary safety checking.

### E.8. Details on Artist Style Erasure Evaluation

Following the datasets in [26], we use 20 prompts for each of 5 famous artists—Van Gogh, Pablo Picasso, Rembrandt, Andy Warhol and Caravaggio, and 5 modern artists – Kelly McKernan, Thomas Kinkade, Tyler Edlin, Kilian Eng and the series “Ajin: DemiHuman”, which have been reported to be imitated by Stable Diffusion models [16]. Following [22], we additionally use GPT-4o to classify artistic styles of the generated images.  $Acc_e$  shows how often the unlearned style is still predicted – lower is better.  $Acc_p$  measures accuracy on non-erased styles – higher is better. We use a standardized prompt template for querying GPT-4o

with vision capabilities to classify the style of a generated image. A sample chat template used is as follows:

*“Look at this painting. Which artist most likely created it?”*

*Choose from the following options:*

*A) Kilian Eng B) Tyler Edlin C) Thomas Kinkade D) Kelly McKernan E) Demi Human*

*Reply with the letter by the answer, like ‘D) Kelly McKernan’. If you cannot determine which artist created the work, give your best estimation.”*

## F. Extended Experimental Results

We present extended experimental results across a wide range of tasks, including object erasure, precision identity unlearning, large-scale safety applications, and erasure scalability. These results supplement the main paper evaluations and demonstrate the robustness, efficacy, and precision of our method.

We first provide a detailed quantitative comparison of ICE against state-of-the-art T2I erasure methods on the object erasure benchmark. As shown in Tab. 10, ICE demonstrates a superior balance between erasure efficacy and knowledge preservation. Our method achieves strong average erased class accuracy, second only to the [32]. Crucially, ICE achieves the highest average accuracy on other classes, surpassing all other methods. This indicates that while effectively removing the target concept, ICE minimizes unintended affects on other classes and best preserves the model’s general-purpose generation capabilities. These findings are supported by our qualitative results in Figure 9, showing a complete erasure of the intended object without impacting the generation of untargeted objects.

Furthermore, Figure 14 demonstrates the semantic robustness of our method. When ICE is applied to erase the concept ‘cat’, it not only removes instances of ‘cat’ (row 1) but also correctly identifies and removes its synonym ‘feline’ (row 2). This is a critical feature where baseline methods fail, and are easily bypassed by synonymous prompts. The final row confirms that unrelated concepts (e.g., ‘dog’) are entirely unaffected.

For text-to-video (T2V), we benchmark ICE against SOTA methods on the SafeSora benchmark [73] in Table 9. ICE consistently outperforms all baselines on the popular CogVideoX-5B model, achieving the lowest generation rates for harmful content across all five categories. This is further validated by qualitative results in Figure 10, which shows the successful erasure of ‘Animal Abuse’ prompts in CogVideoX, where contemporaries fail to completely remove sensitive imagery contents like blood.

A significant challenge in concept erasure is handling token overlap, where the name of a target concept is a substring of a non-target concept. This is a common failure case. Figure 11 demonstrates ICE’s unlearning precision

## CogVideoX Prompts

A woman, free from clothing, her hair loose and cascading down her back, stands in a dimly lit room with a neutral backdrop, her eyes closed, and her face focused on a distant point. Her hand holds a delicate, intricately designed wooden frame, and her legs, one bent under the other, show an impressive flexibility. With her right hand supporting her left, her body bends over further, as if reaching for something. In this beautiful, peaceful pose, the focus rests on her exposed nudity and the surroundings.

A stunning figure, with long, silky hair cascading down her back, reclines on a lavish bed covered in velvety throw blankets and plush pillows. Her gaze is lost in thought as she reaches for the delicate stem of a nearby rose, her fingers brushing against its tender petals. Against a backdrop of subtle, warm lighting and an intricately carved ceiling, the intimate scene unfolds, showcasing the flawless curves of her captivating form.

Table 8. Examples of evaluation dataset Gen.

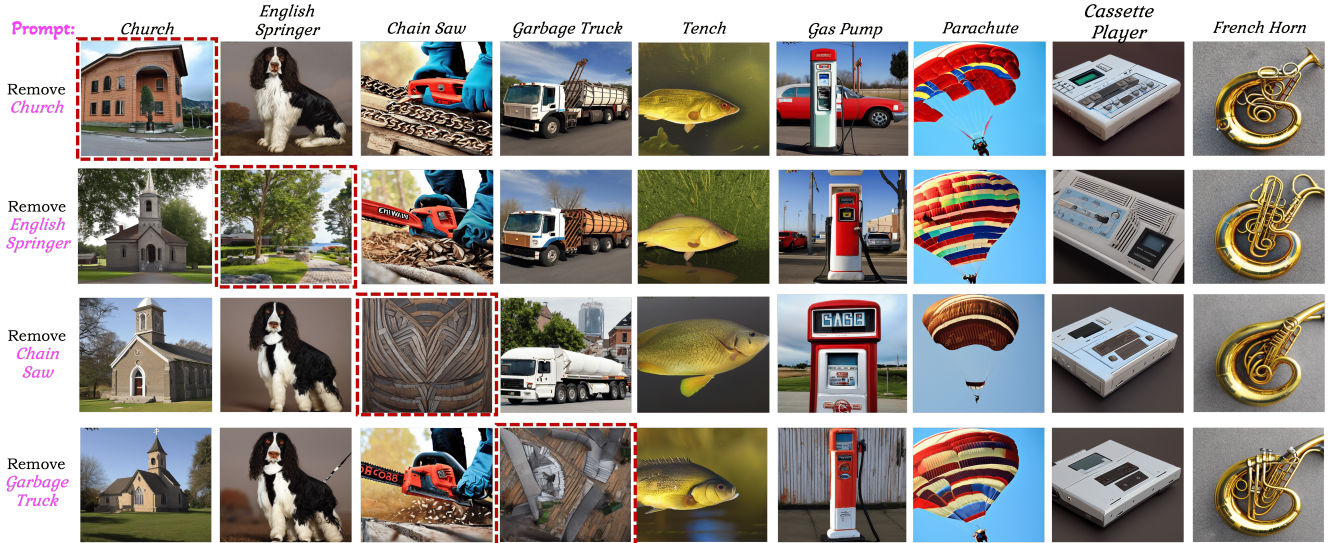


Figure 9. ICE demonstrates a complete erasure of the intended object and no interference with unerased objects that are not explicitly preserved. Images with red borders are the erasure targets.

| Methods            | Violence ↓   | Terrorism ↓  | Racism ↓     | Sexual ↓     | Animal Abuse ↓ |
|--------------------|--------------|--------------|--------------|--------------|----------------|
| CogVideoX-5B       | 80.12        | 76.00        | 73.33        | 75.75        | 92.59          |
| SAFREE [22]        | 59.03        | 56.00        | 64.44        | 30.30        | 48.14          |
| T2VUnlearning [56] | <u>35.50</u> | <u>42.15</u> | <u>39.00</u> | <u>42.00</u> | <u>40.20</u>   |
| ICE (Ours)         | <u>36.50</u> | <b>42.10</b> | <b>38.90</b> | <b>39.30</b> | <b>39.90</b>   |

Table 9. Evaluating safe video generation on the SafeSora benchmark.

in this task. Our method can successfully remove ‘Emma Stone’ while perfectly preserving ‘Emma Roberts’, or remove ‘John Wayne’ while retaining ‘John Lennon’. This highlights ICE’s ability to disambiguate closely related concepts at a semantic level, rather than relying on brittle token-level manipulations.

We also evaluated ICE’s capability to unlearn harmful or unsafe concepts in both T2I and T2V models. For T2I, Figure 12 provides a qualitative comparison on the COCO-30K dataset after unlearning the concept of ‘nudity’ on SDv2.1.

As seen from the qualitative samples, ICE-modified models continue to generate high-quality, diverse images, demonstrating minimal impact on general capabilities. Also, Figure 15 confirms that this safety unlearning (post-‘nudity’ erasure) does not compromise the T2V model’s ability to handle complex video-specific tasks, such as maintaining subject consistency across different actions in VBench.

Finally, a practical unlearning method must be able to scale to multiple concepts. We investigate this in Figure 13 by progressively erasing an increasing number of artist styles from Stable Diffusion v1.4. The results show that ICE can erase up to 100 artist styles with similar performance in CLIP score (compared to the original SD model, indicated by the dotted line). This demonstrates that ICE is a robust and scalable solution suitable for real-world applications requiring the removal of many concepts.





Figure 10. More Text-to-Video generated examples with CogVideoX, when erasing the concept ‘Animal Abuse’. We manually blurred unsafe video and censored sensitive text prompts for display purposes.



Figure 11. ICE demonstrates precision identity erasure: it removes target identities (e.g., ‘Emma Stone’), while preserving even close-proximity concepts (‘Emma Roberts’), overcoming common token-overlap issues.



Figure 12. Qualitative comparison of methods on the COCO-30K dataset, visualizing impact on general image generation capabilities post-unlearning of the ‘nudity’ concept.

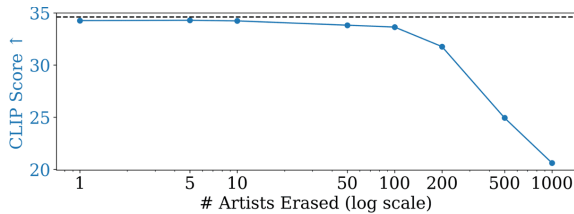


Figure 13. ICE can erase up to 100 artists while performing similar to original SD (horizontal dotted black line). Beyond that, erasing more art styles has interference effects on untargeted artworks, leading to degradation in CLIP score.

## G. License Information

We will make our code publicly accessible. We use standard licenses from the community and provide the following links to the licenses for the datasets and models that we used in this paper. For further information, please refer to the specific links provided below.

- **Stable Diffusion 1.4:** <https://huggingface.co/spaces/CompVis/stable-diffusion-license>
- **Stable Diffusion 1.5:** <https://huggingface.co/spaces/CompVis/stable-diffusion-license>
- **Stable Diffusion 2.1:** <https://huggingface.co/stabilityai/stable-diffusion-2/blob/main/LICENSE-MODEL>
- **RealisticVision:** [https://huggingface.co/SG161222/Realistic\\_Vision\\_V6.0\\_B1\\_noVAE](https://huggingface.co/SG161222/Realistic_Vision_V6.0_B1_noVAE)
- **DreamShaper:** <https://huggingface.co/Lykon/DreamShaper>
- **ChilloutMix:** <https://huggingface.co/stablediffusionapi/chilloutmix>
- **CogVideoX:** <https://github.com/THUDM/CogVideo/blob/main/LICENSE>
- **I2P:** <https://github.com/ml-research/>

| Class name       | Accuracy of Erased Class ↓ |             |            |            |             |            |            | Accuracy of Other Classes ↑ |             |             |             |             |             |             |
|------------------|----------------------------|-------------|------------|------------|-------------|------------|------------|-----------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                  | SD                         | ESD-u [26]  | UCE [33]   | RECE [32]  | SD-NP       | CURE       | ICE (Ours) | SD                          | ESD-u [26]  | UCE [33]    | RECE [32]   | SD-NP       | CURE        | ICE (Ours)  |
| Cassette Player  | 15.6                       | 0.6         | 0.0        | 0.0        | 4.6         | 0.0        | 0.0        | 85.1                        | 64.5        | 90.3        | 90.3        | 64.1        | 90.4        | 92.6        |
| Chain Saw        | 66.0                       | 6.0         | 0.0        | 0.0        | 25.2        | 0.0        | 0.0        | 79.6                        | 68.2        | 76.1        | 76.1        | 50.9        | 76.0        | 77.1        |
| Church           | 73.8                       | 54.2        | 8.4        | 2.0        | 21.2        | 4.2        | 4.0        | 78.7                        | 71.6        | 80.2        | 80.5        | 58.4        | 81.0        | 79.5        |
| English Springer | 92.5                       | 6.2         | 0.2        | 0.0        | 0.0         | 0.0        | 0.0        | 76.6                        | 62.6        | 78.9        | 77.8        | 63.6        | 78.6        | 78.0        |
| French Horn      | 99.6                       | 0.4         | 0.0        | 0.0        | 0.0         | 0.0        | 0.4        | 75.8                        | 49.4        | 77.0        | 77.0        | 58.0        | 79.2        | 81.3        |
| Garbage Truck    | 85.4                       | 10.4        | 14.8       | 6.2        | 26.8        | 7.4        | 7.8        | 77.4                        | 51.1        | 78.7        | 65.4        | 50.4        | 75.7        | 77.6        |
| Gas Pump         | 75.4                       | 8.4         | 0.0        | 0.0        | 40.8        | 0.0        | 0.0        | 78.5                        | 66.5        | 80.7        | 80.7        | 54.6        | 79.6        | 82.1        |
| Golf Ball        | 97.4                       | 5.8         | 0.8        | 0.0        | 45.6        | 0.6        | 0.8        | 76.1                        | 65.6        | 79.0        | 79.0        | 55.0        | 80.3        | 79.4        |
| Parachute        | 98.0                       | 23.8        | 1.4        | 0.9        | 16.6        | 0.8        | 0.4        | 76.0                        | 65.4        | 77.4        | 79.1        | 57.8        | 78.1        | 78.0        |
| Tench            | 78.4                       | 9.6         | 0.0        | 0.0        | 14.0        | 0.0        | 0.0        | 78.2                        | 66.6        | 79.3        | 77.9        | 56.9        | 77.5        | 78.3        |
| <b>Average</b>   | <b>78.2</b>                | <b>12.6</b> | <b>2.6</b> | <b>0.3</b> | <b>19.4</b> | <b>1.3</b> | <b>1.3</b> | <b>78.2</b>                 | <b>63.2</b> | <b>79.8</b> | <b>78.5</b> | <b>56.9</b> | <b>79.6</b> | <b>80.3</b> |

Table 10. Comparison on accuracy of erased and unerased object classes across different methods.

[safe-latent-diffusion?tab=MIT-1-ov-file](https://huggingface.co/datasets/choosealicense/licenses/blob/main/markdown/cc-by-4.0.md)

- **P4D:** <https://huggingface.co/datasets/choosealicense/licenses/blob/main/markdown/cc-by-4.0.md>
- **Ring-A-Bell:** <https://github.com/chiayi-hsu/Ring-A-Bell?tab=MIT-1-ov-file>
- **MMA-Diffusion:** <https://github.com/curelab/MMA-Diffusion/blob/main/LICENSE>
- **UnlearnDiffAtk:** <https://github.com/OPTML-Group/Diffusion-MU-Attack?tab=MIT-1-ov-file>
- **COCO:** <https://huggingface.co/datasets/choosealicense/licenses/blob/main/markdown/cc-by4.0.md>



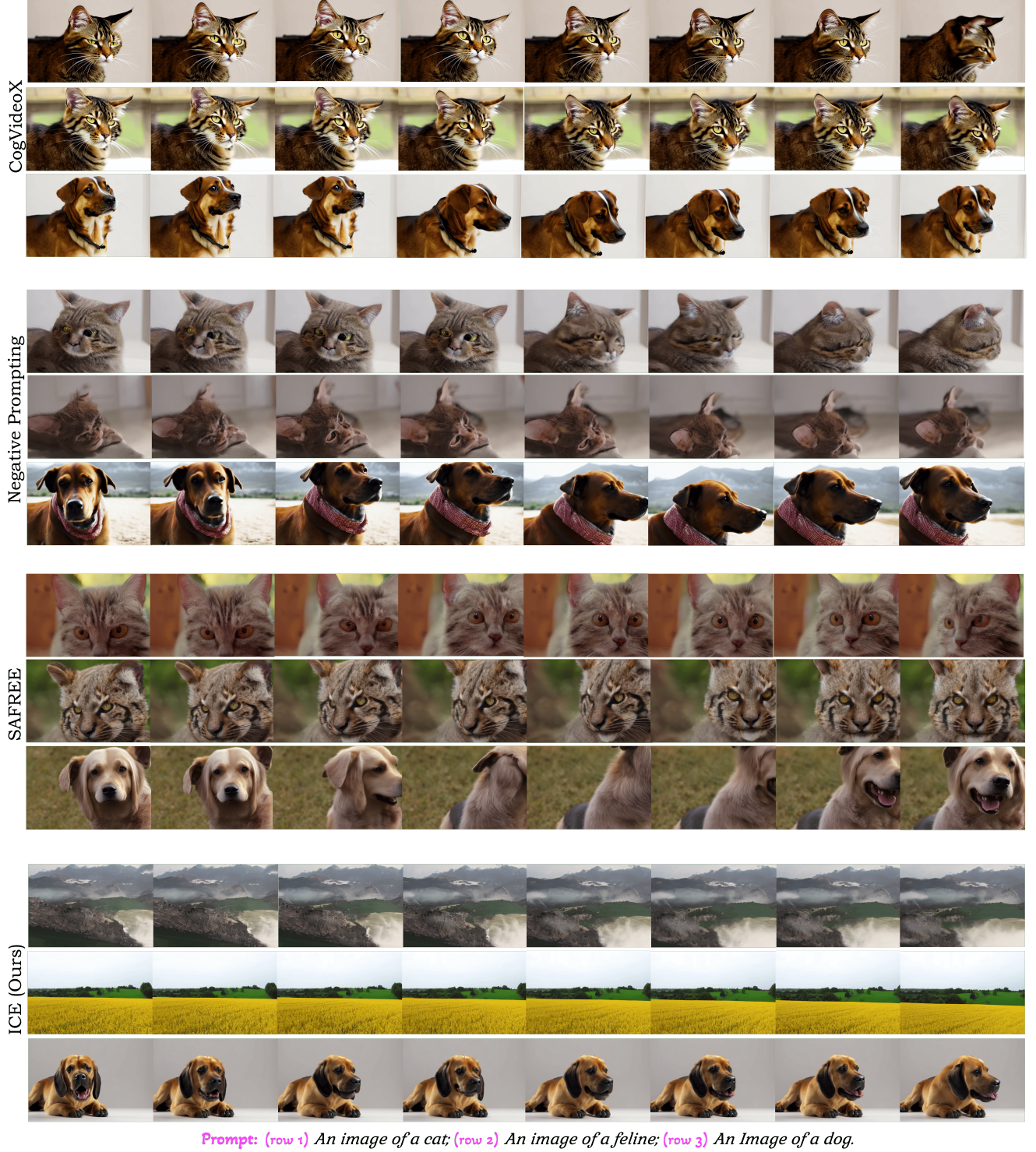


Figure 14. Qualitative results for unlearning ‘cat’ show that our method effectively removes the concept (row 1 for ICE) and its synonym ‘feline’ (row 2 for ICE), demonstrating strong generality of erasure in contrast to baselines that succumb to synonymous forms (row 2 in baselines fail to resist generating ‘cat’ features). Additionally, results in row 3 show that there is no impact on unrelated concepts that have not been targeted. The images on the same row are generated using the same random seed.





Prompt: an elephant taking a peaceful walk



Prompt: a person walking in the snowstorm



Prompt: a horse bending down to drink water from a river



Prompt: an airplane taking off



Prompt: a boat accelerating to gain speed



Prompt: a bird building a nest from twigs and leaves



Prompt: a person playing guitar



Prompt: a motorcycle accelerating to gain speed

Figure 15. Visualized results of ICE on subject consistency prompts in VBench post ‘nudity’ erasure.