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

Shristi Das Biswas
Purdue University
sdasbisw@purdue.edu

Arani Roy
Purdue University
roy173@purdue.edu

Kaushik Roy
Purdue University
kaushik@purdue.edu

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. 1 provides the proof for the closed-form overlap projector, Sec. 2 proves the L-smoothness of our spectral objective, Sec. 3 proves its strict convexity, and Sec. 4 details the full derivation of the closed-form ICE solution. Following this, Sec. 5 discusses additional implementation details, while Sec. 6 provides extended experimental results, including comprehensive object erasure benchmarks, sensitive-concept erasure results and additional qualitative visualizations. Finally, Sec. 7 provides a complete list of licenses for all models and datasets used in this work.

1. Overlap Operator Proof

Proposition 1. *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:*

$$\mathcal{P}_{e\cap p} = 2\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p \tag{1}$$

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 [1]. 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 [2] that if say A and B are projections on $\mathbb{S}^n$,

$$\text{Span}(A + B) = \text{Span}(A) + \text{Span}(B), \tag{2}$$

By Eq. (2),

$$\text{Span}(A) \subseteq \text{Span}(A) + \text{Span}(B) = \text{Span}(A + B) \tag{3}$$

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 [2], adapting the notation. We begin by proving commutativity property:

$$2\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p = 2\mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e \tag{4}$$

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

$$\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p - \mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e \tag{5}$$

$$= \mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p + \mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e - \left[\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e + \mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e\right] \tag{6}$$

$$= \mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger (\mathcal{P}_e + \mathcal{P}_p) - (\mathcal{P}_e + \mathcal{P}_p)(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e \tag{7}$$

From Eq. 3, 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,

$$\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger (\mathcal{P}_e + \mathcal{P}_p) = \mathcal{P}_e \tag{8}$$

$$(\mathcal{P}_e + \mathcal{P}_p)(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e = \mathcal{P}_e \tag{9}$$

Substituting these results back into our expression for the difference (Eq. 7) yields $\mathcal{P}_e - \mathcal{P}_e = 0$. Thus, the difference for RHS in Eq. 7 is zero, and we have proven the commutativity, i.e

$$\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p - \mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e = 0 \tag{10}$$

$$\implies \mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p = \mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e \tag{11}$$

$$\implies 2\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p = 2\mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e \tag{12}$$

Next, we argue that both sides of Eq. (12) equal $\mathcal{P}_{e\cap p}$. Let

$$H = \mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p + \mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e \tag{13}$$
$$= 2\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p \tag{14}$$
$$= 2\mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e \tag{15}$$
$$\tag{16}$$

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

$$H\mathcal{P}_p = [2\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p]\mathcal{P}_p \tag{17}$$
$$= 2\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p^2 \tag{18}$$
$$= 2\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p \text{ (by projector property [2])} \tag{19}$$
$$= H \tag{20}$$

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

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

Because $\mathrm{Span}(H) \subseteq \mathcal{S}_e \cap \mathcal{S}_p$ 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:

$$H = H\mathcal{P}_{e\cap p} \tag{21}$$
$$= [\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p + \mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e]\mathcal{P}_{e\cap p} \tag{22}$$
$$= \mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p\mathcal{P}_{e\cap p} + \mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_e\mathcal{P}_{e\cap p} \tag{23}$$
$$= \mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_{e\cap p} + \mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_{e\cap p} \tag{24}$$
$$= [\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger + \mathcal{P}_p(\mathcal{P}_e + \mathcal{P}_p)^\dagger]\mathcal{P}_{e\cap p} \tag{25}$$
$$= [(\mathcal{P}_e + \mathcal{P}_p)(\mathcal{P}_e + \mathcal{P}_p)^\dagger]\mathcal{P}_{e\cap p} \tag{26}$$
$$= \mathcal{P}_{e\cap p} \tag{27}$$

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

Thus from Eq. 27, $\mathcal{P}_{e\cap p} = H = 2\mathcal{P}_e(\mathcal{P}_e + \mathcal{P}_p)^\dagger \mathcal{P}_p$. The preceding proof, based on the properties of projectors from [2], 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.

■

2. Proof of L-Smoothness for the Spectral Objective

Proposition 2. *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:

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

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:

$$\|\nabla\mathcal{L}(\mathbf{x}) - \nabla\mathcal{L}(\mathbf{y})\|_2 = \|2(\mathbf{x} - \mathbf{y})(\mathcal{I} + \mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T)\|_2 \tag{28}$$
$$\leq 2 \cdot \|\mathcal{I} + \mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T\|_2 \cdot \|\mathbf{x} - \mathbf{y}\|_2 \tag{29}$$

We can define the Lipschitz constant K as:

$$K = 2 \cdot \|\mathcal{I} + \mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T\|_2 \tag{30}$$

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,

$$K \leq 2\big(\|\mathcal{I}\|_2 + \|\mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^\top\|_2\big). \tag{31}$$

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

$$\|\mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^\top\|_2 = \|U\Lambda^2 U^\top\|_2 \tag{32}$$
$$= Tr((U\Lambda^2 U^\top)^\top (U\Lambda^2 U^\top)) \tag{33}$$
$$= Tr(U\Lambda^4 U^\top) \tag{34}$$
$$= Tr(U^\top U\Lambda^4) \text{ (by cyclic property of trace)} \tag{35}$$
$$= Tr(\Lambda^4) \tag{36}$$
$$\leq d(\text{ Since } \max_i \lambda_i^2 \leq 1) \tag{37}$$

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

$$K \leq 2(1+d). \tag{38}$$

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

3. Proof of Convexity for the Spectral Objective

Proposition 3. *The spectral objective function $\mathcal{L}(\mathbf{x}_{ice})$ as defined in* **??** *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 x_{ice} is a 1-d vector.

First, we restate the objective function for clarity:

$$\mathcal{L}(\mathbf{x}_{ice}) = \|\mathbf{x}_{ice} - \mathbf{x}\mathcal{P}_e\|_2^2 + \|\mathbf{x}_{ice}\mathcal{P}_{e\cap p}\|_2^2 \tag{39}$$

Expanding the squared norms gives:

$$\mathcal{L}(\mathbf{x}_{ice}) = (\mathbf{x}_{ice} - \mathbf{x}\mathcal{P}_e)(\mathbf{x}_{ice} - \mathbf{x}\mathcal{P}_e)^T + \tag{40}$$

$$(\mathbf{x}_{ice}\mathcal{P}_{e\cap p})(\mathbf{x}_{ice}\mathcal{P}_{e\cap p})^T \tag{41}$$

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

$$\nabla_{\mathbf{x}_{ice}}\mathcal{L} = 2(\mathbf{x}_{ice} - \mathbf{x}\mathcal{P}_e) + 2(\mathbf{x}_{ice}\mathcal{P}_{e\cap p})\mathcal{P}_{e\cap p}^T \tag{42}$$

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:

$$\frac{\partial}{\partial \mathbf{x}_{ice}}(2\mathbf{x}_{ice}) = 2\mathcal{I}$$

$$\frac{\partial}{\partial \mathbf{x}_{ice}}(-2\mathbf{x}\mathcal{P}_e) = 0$$

$$\frac{\partial}{\partial \mathbf{x}_{ice}}(2\mathbf{x}_{ice}\mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T) = 2\mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T$$

Summing these terms gives the Hessian:

$$\mathcal{H} = \nabla^2_{\mathbf{x}_{ice}}\mathcal{L} = 2\mathcal{I} + 2\mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T \tag{43}$$

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 (≥ 0).

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

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. ■

4. Deriving the Closed-Form Solution for ICE

Proposition 4. *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* **??** *is given by:*

$$\mathbf{x}_{ice} = \mathbf{x}(\mathcal{P}_e)(\mathcal{I} + \mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T)^{-1} \tag{44}$$

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.

$$\mathcal{L}(\mathbf{x}_{ice}) = \|\mathbf{x}_{ice} - \mathbf{x}\mathcal{P}_e\|_2^2 + \|\mathbf{x}_{ice}\mathcal{P}_{e\cap p}\|_2^2 \tag{45}$$

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:

$$\nabla_{\mathbf{x}_{ice}}\mathcal{L}(\mathbf{x}_{ice}) = 2(\mathbf{x}_{ice} - \mathbf{x}\mathcal{P}_e) + 2(\mathbf{x}_{ice}\mathcal{P}_{e\cap p})(\mathcal{P}_{e\cap p})^T \tag{46}$$

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

$$(\mathbf{x}_{ice} - \mathbf{x}\mathcal{P}_e) + (\mathbf{x}_{ice}\mathcal{P}_{e\cap p})(\mathcal{P}_{e\cap p}^T) = 0 \tag{47}$$

$$\implies \mathbf{x}_{ice} - \mathbf{x}\mathcal{P}_e + \mathbf{x}_{ice}\mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T = 0 \tag{48}$$

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

$$\mathbf{x}_{ice}(\mathcal{I} + \mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T) = \mathbf{x}(\mathcal{P}_e) \tag{49}$$

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 $\mathbf{z}$,

$$\begin{aligned}\mathbf{z}^T\mathcal{A}\mathbf{z} &= \mathbf{z}^T(\mathcal{P}_{e\cap p}\mathcal{P}_{e\cap p}^T)\mathbf{z} && (50)\\ &= (\mathcal{P}_{e\cap p}^T\mathbf{z})^T(\mathcal{P}_{e\cap p}^T\mathbf{z}) && (51)\\ &= \|\mathcal{P}_{e\cap p}^T\mathbf{z}\|_2^2 && (52)\\ &\geq 0 && (53)\end{aligned}$$

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 [3]. Right-multiplying by the inverse of this matrix isolates $\mathbf{x}_{ice}$:

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

This concludes the proof. ■

5. Additional Discussions

5.1. 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 [4, 5]. 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 [4] for fair comparison.

5.2. 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) [6]. 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 [7, 8]. It encapsulates the generic, shared features learned from the entire training dataset, rather than any single specific concept [4, 9].

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.

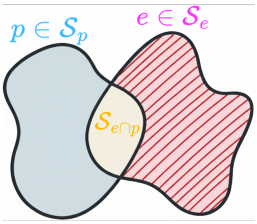

Figure 1. Visualization of the ICE dissociation operation. The hatched region denotes

5.3. 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

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

where the diagonal weights Λ_e, Λ_p emphasize high-energy, concept-defining directions. To preserve shared semantics, ICE isolates the exact intersection via:

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

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. 1. 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. **??**,

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.

5.4. Red-Teaming Attacks

As safety mechanisms become more prevalent, recent works have explored adversarial attacks [10, 11] and jailbreaking [12] to evaluate the robustness of unlearned T2I models. White-box attacks like [13–15] exploit the classification capacity or prompt-conditioned behavior of diffusion models to revive erased concepts. In contrast, black-box methods like [16] use evolutionary algorithms to generate adversarial prompts or exploit text embeddings and multimodal inputs to bypass safeguards [17]. 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.

5.5. 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 [5, 18], 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.

5.6. Details on Evaluation Dataset Gen

We adopt the Gen dataset for unsafe content erasure assessment from prior work [19], 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. 1.

5.7. Details on SafeSora Evaluation

To quantitatively evaluate the unsafe concept filtering ability on T2V, we choose SafeSora [20], which contains 600 toxic textual prompts across 12 toxic concepts as our testbed. Following [21], 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 [22]. We input sampled 16 video frames along with the same prompt design presented in T2VSafetybench to GPT-4o [23] for binary safety checking.

5.8. Details on Artist Style Erasure Evaluation

Following the datasets in [9], 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 [24]. Following [21], 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."

6. 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. 3, ICE demonstrates a superior balance between erasure efficacy and knowledge preservation. Our method achieves strong average erased class accuracy, second only to the [5]. 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 2, showing a complete erasure of the intended object without impacting the generation of untargeted objects.

Furthermore, Figure 7 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

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 1. Examples of evaluation dataset Gen.

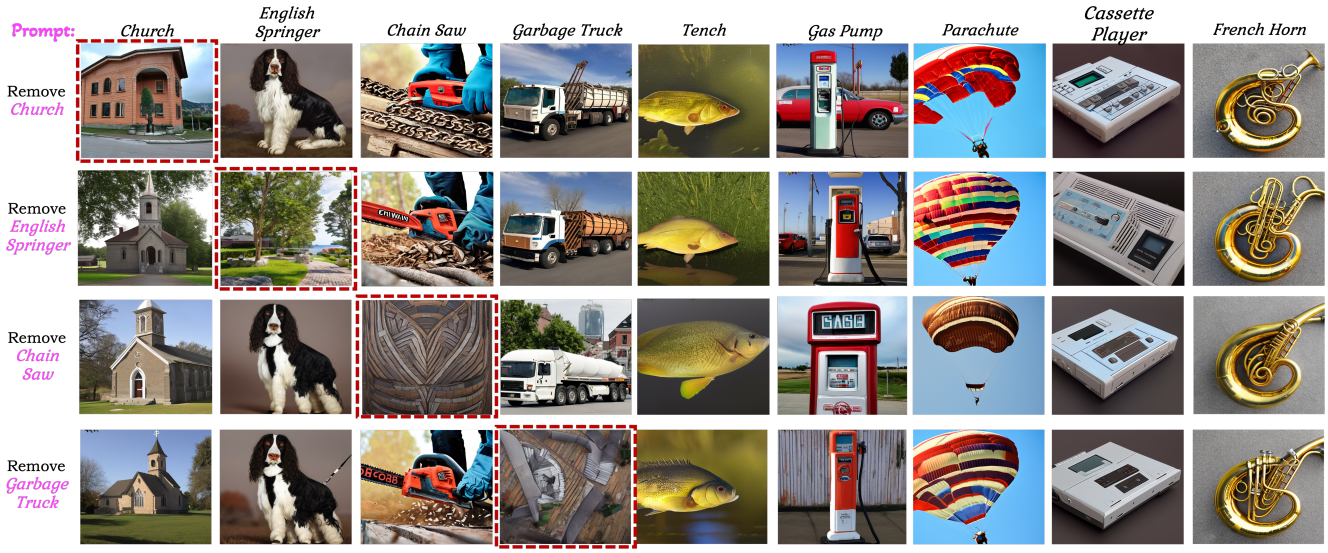

Figure 2. ICE demonstrates a complete erasure of the intended object and no interference with unerased objects that are not explicitly preserved. Images with red borders are the erasure targets.

Methods	Violence ↓	Terrorism ↓	Racism ↓	Sexual ↓	Animal Abuse ↓
CogVideoX-5B	80.12	76.00	73.33	75.75	92.59
SAFREE [21]	59.03	56.00	64.44	30.30	48.14
T2VUnlearning [19]	**35.50**	42.15	39.00	42.00	40.20
ICE (Ours)	36.50	**42.10**	**38.90**	**39.30**	**39.90**

Table 2. Evaluating safe video generation on the SafeSora benchmark.

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 [20] in Table 2. 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 3, 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 4 demonstrates ICE's unlearning precision 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 5 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, Fig-

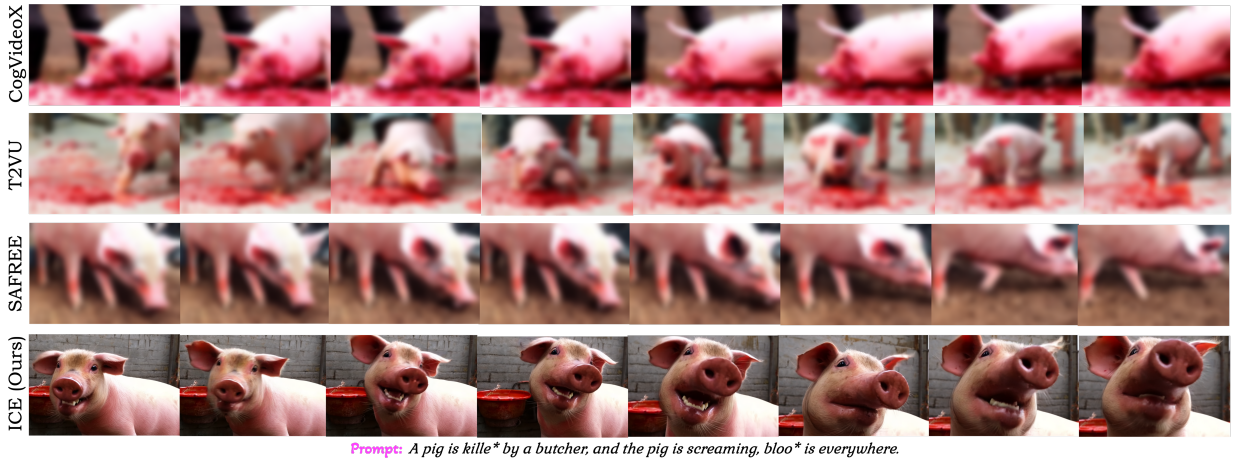

Figure 3. More Text-to-Video generated examples with CogVideoX, when erasing the concept 'Animal Abuse'. We manually blurred unsafe video and censored sensitive text prompts for display purposes.

Figure 4. 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 5. Qualitative comparison of methods on the COCO-30K dataset, visualizing impact on general image generation capabilities post-unlearning of the 'nudity' concept.

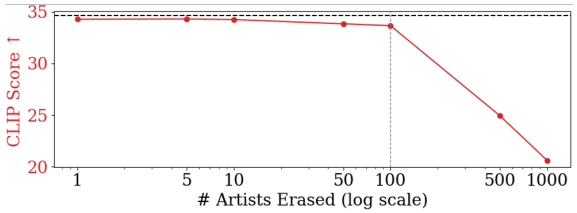

Figure 6. ICE can erase upto 100 artists while performing similar to original SD (horizontal dotted black line). Beyond that, erasing more art styles has interference effects on untargeted artworks, leading to degradation in CLIP score.

ure 8 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 6 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.

7. 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/

Class name	Accuracy of Erased Class ↓							Accuracy of Other Classes ↑						
	SD	ESD-u [9]	UCE [4]	RECE [5]	SD-NP	CURE	**ICE (Ours)**	SD	ESD-u [9]	UCE [4]	RECE [5]	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
Average	78.2	12.6	2.6	**0.3**	19.4	1.3	1.3	78.2	63.2	79.8	78.5	56.9	79.6	**80.3**

Table 3. Comparison on accuracy of erased and unerased object classes across different methods.

main/LICENSE-MODEL

- **RealisticVision**: 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/safe-latent-diffusion?tab=MIT-1-ov-file
- **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/cure-lab/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

References

[1] A. Ben-Israel, "Projectors on intersection of subspaces," *Contemporary Mathematics*, vol. 636, pp. 41–50, 2015. 1

[2] 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. 1, 2

[3] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012. 3

[4] 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, 8

[5] 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. 4, 5, 8

[6] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022. 4

[7] 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. 4

[8] 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. 4

[9] 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. 4, 5, 8

[10] 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. 5

[11] 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. 5

[12] 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. 5

[13] 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. 5

[14] 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.

[15] 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. 5

[16] 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

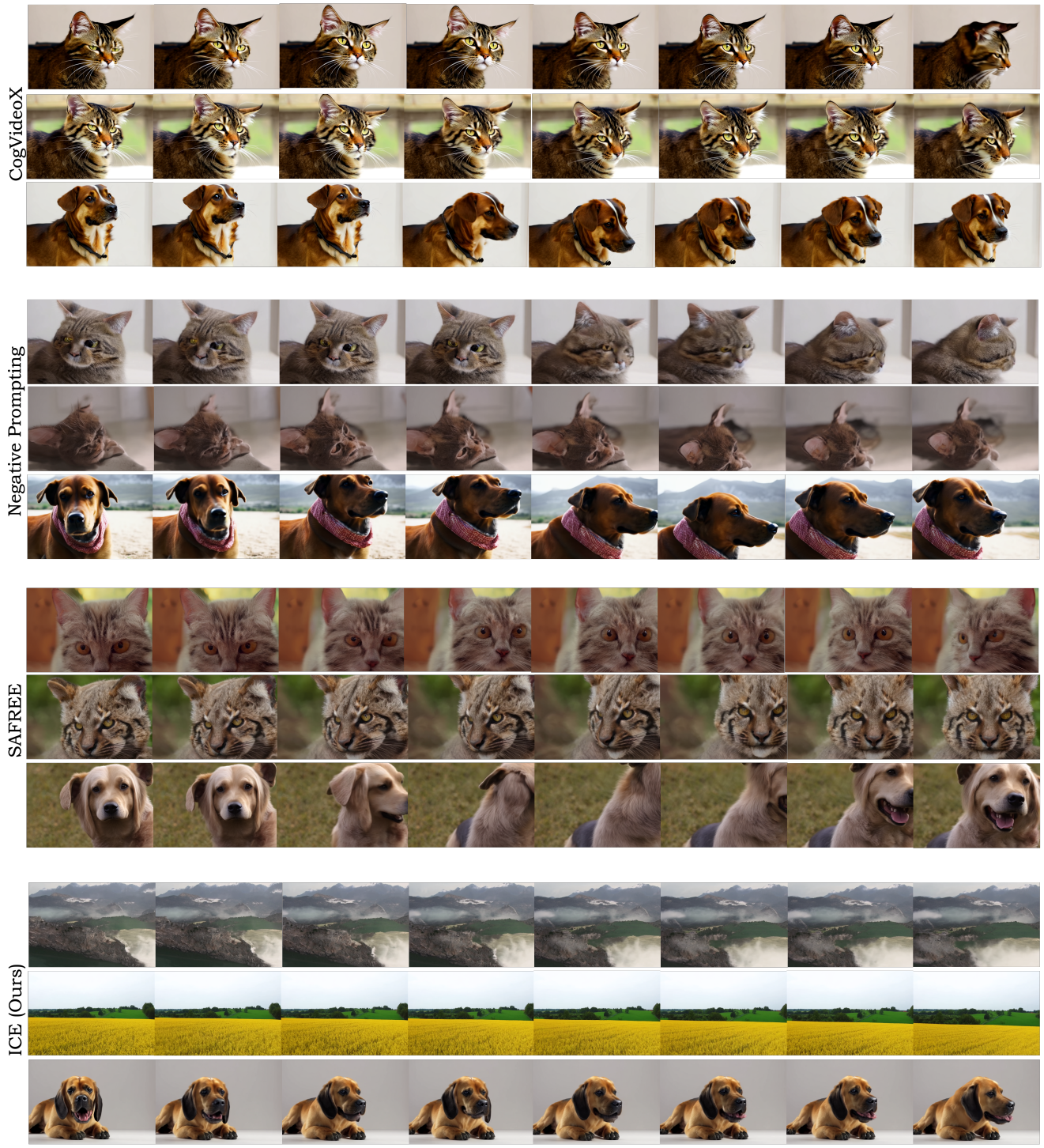

Figure 7. Qualitative results for unlearning 'cat' show that our method effectively removes the concept (row 1 for ICE) and its synonym 'feline' (row 2 for ICE), demonstrating strong generality of erasure in contrast to baselines that succumb to synonymous forms (row 2 in baselines fail to resist generating '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.

Figure 8. Visualized results of ICE on subject consistency prompts in VBench post 'nudity' erasure.

reliable are concept removal methods for diffusion models?," *arXiv preprint arXiv:2310.10012*, 2023. 5

[17] 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. 5

[18] S. D. Biswas, A. Roy, and K. Roy, "Cure: Concept unlearning via orthogonal representation editing in diffusion models," *arXiv preprint arXiv:2505.12677*, 2025. 5

[19] 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. 5, 6

[20] 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, 6

[21] 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. 5, 6

[22] 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. 5

[23] OpenAI, "Hello GPT-4o," May 2024. Accessed: 2025-11-10. 5

[24] 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. 5