ETCE: Efficient Two-Stage Concept Erasure for Text-to-Image Diffusion Models - Supplementary Materials

A Derivation

A.1 Derivation of the Closed-Form Solution for ETCE

Given the stacked embedding vectors of non-target (c_p), target (c_e), and anchor (c_a) concepts, and letting W_1 denote the original Stable Diffusion value matrices, we derive the solution through the following equation, where W_2 represents the post-editing Stable Diffusion value matrices:

$$min||W_2 c_e - W_1 c_a||^2 + ||W_2 - W_1||^2, s.t., W_2 c_p - W_1 c_p = 0. \quad (1)$$

We define $W_2 - W_1 = \Delta P$ such that Equation 1 can be expressed as:

$$min||(W_1 + \Delta P)c_e - W_1 c_a||^2 + ||\Delta P||^2, s.t., (\Delta P)c_p = 0, \quad (2)$$

where P is a projection matrix constructed from c_p that maps the update Δ onto this null space.

The solution is obtained via Lagrange multipliers, beginning with the formulation of this target function:

$$\mathcal{L}(\Delta, \gamma) = ||(W_1 + \Delta P)c_e - W_1 c_a||^2 + ||\Delta P||^2 + \gamma^\top \left((\Delta P)c_p\right). \quad (3)$$

Setting the gradient of the $\mathcal{L}(\Delta, \gamma)$ with respect to Δ to zero yields the optimality condition for W_2:

$$\frac{\partial \mathcal{L}(\Delta, \gamma)}{\partial \Delta} = 2\left((W_1 + \Delta P)c_e - W_1 c_a\right) c_e^\top P^\top + 2\Delta P P^\top + \gamma c_p^\top P^\top = 0. \quad (4)$$

The closed-form solution for the parameter update matrix ΔP can be directly derived from Equation 4:

$$\Delta P = \left(W_1 c_a c_e^\top P - W_1 c_e c_e^\top P - \frac{1}{2}\gamma c_p^\top P\right)\left(c_e c_e^\top P + I\right)^{-1}. \quad (5)$$

We define $M = (c_e c_e^\top P + I)^{-1}$ to simplify Equation 5. By substituting the simplified expression into the constraint condition in Equation 2, we obtain:

$$\left(W_1 c_a c_e^\top P - W_1 c_e c_e^\top P - \frac{1}{2}\gamma c_p^\top P\right) M c_p = 0. \quad (6)$$

Based on this equation, we obtain:

$$\gamma = 2W_1(c_a c_e^\top - c_e c_e^\top) P M c_p (c_p^\top P M c_p)^{-1}. \quad (7)$$

By substituting the value from Equation 7 into Equation 5, we obtain:

$$\Delta P = W_1\left(c_a c_e^\top - c_e c_e^\top\right) PQM, \quad (8)$$

where $Q = I - Mc_p\left(c_p^\top P M c_p\right)^{-1} c_p^\top P$, and $M = \left(c_e c_e^\top P + I\right)^{-1}$. Therefore, the solution yields:

$$W_2 = W_1\left(I + \left(c_a c_e^\top - c_e c_e^\top\right) PQM\right). \quad (9)$$

A.2 Derivation of the Optimal Target Embedding c_e'

Given the derived matrices W_2, W_1, and target embedding c_e, we formulate and solve the following optimization problem:

$$\min \left\|\mathrm{W}_2 c_e' - \mathrm{W}_1 c_e\right\|^2 + \alpha \left\|\mathrm{c}_\mathrm{e}'\right\|^2. \quad (10)$$

Here, we first derive the objective function as follows:

$$L(c_e') = \left\|\mathrm{W}_2 c_e' - \mathrm{W}_1 c_e\right\|^2 + \alpha \left\|\mathrm{c}_\mathrm{e}'\right\|^2. \quad (11)$$

Subsequently, we set the derivative of the objective function L with respect to c_e' to zero:

$$\frac{dL}{dc_e'} = 2(W_2^\top W_2 c_e' - W_2^\top W_1 c_e) + 2\alpha c_e' = 0. \quad (12)$$

Solving equation 12 leads to the solution of the optimal target embedding:

$$c_e' = \left(\alpha I + W_2^T W_2\right)^{-1}\left(W_2^T W_1\right) c_e. \quad (13)$$

B MORE QUANTITATIVE RESULTS

Due to the page limit in the main paper, we include additional experiments here to provide a comprehensive evaluation of ETCE's performance.

We evaluate the algorithm's performance across multiple concept erasure tasks, including style erasure and NSFW content removal. For each task, we design 30 distinct templates to assess its effectiveness. With the adjustable parameters set to $\alpha = 0.001$ and $\beta = 0.003$, we continue to employ the CS and FID metrics to assess the algorithm's capability in erasing the target concepts while preserving non-target concepts.

Table 1: Quantitative comparison in other concept erasure tasks.

Task	**Style Erasure**		**NSFW Content Erasure**	
Concept	Monet	Picasso	Naked	Dressed
	CS ↓	CS ↓	CS ↓	CS ↓
	Erase **Monet**		Erase **Naked**	
SD v1.4	28.91	27.98	26.76	24.82
	CS ↓	FID ↓	CS ↓	FID ↓
ESD	27.01	66.01	25.77	47.31
UCE	25.49	71.13	25.37	45.18
SPEED	24.99	39.05	26.22	32.08
Ours	**24.10**	**37.41**	**25.19**	**28.35**

As can be seen from Table 1, our algorithm achieves the lowest CS and FID values across multiple tasks, further demonstrating its exceptional capability in target concept erasure and non-target concept preservation.

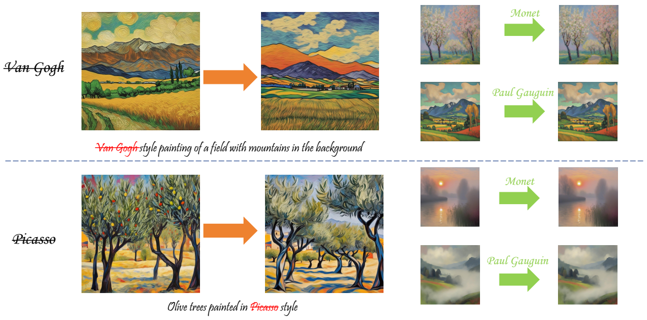

Figure 2: Style erasure of ETCE. Our method effectively preserves image content while removing specific artistic styles, without affecting other author-specific style concepts.

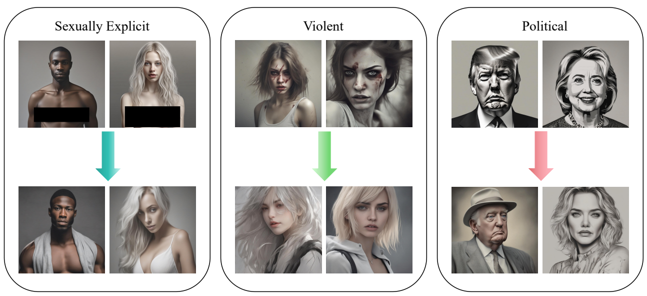

Figure 3: Application of ETCE in Not-Safe-For-Work (NSFW) Content Erasure. Our algorithm can be extended to remove NSFW content, effectively preventing the generation of inappropriate content (e.g., sexual, violent, or political elements).

Additionally, we validated the transferability of our algorithm on SDXL. Here, we set the adjustable parameters to $\alpha = 0.001$ and $\beta = 2$. The performance was evaluated using 80 distinct templates from the main paper for instance erasure, along with 30 templates each for style erasure and NSFW content erasure as introduced earlier. It is worth noting that, since no updated versions of ESD and UCE are available, comparative experiments were only conducted between SPEED and our proposed method.

Table 2: Quantitative Evaluation of Various Concept Erasure Tasks on SDXL

Task	**Instance Erasure**		**Style Erasure**		**NSFW Content Erasure**	
Concept	Mickey	Totoro	Monet	Picasso	Naked	Dressed
	CS ↓	CS ↓	CS ↓	CS ↓	CS ↓	CS ↓
	Erase **Mickey**		Erase **Monet**		Erase **Naked**	
SDXL	26.84	27.62	28.29	28.01	27.25	26.14
	CS ↓	FID ↓	CS ↓	FID ↓	CS ↓	FID ↓
SPEED	27.03	26.23	25.43	31.61	27.16	16.27
Ours	**21.75**	**16.83**	**24.74**	**28.02**	**24.14**	**15.89**

As shown in Table 2, our algorithm also achieves impressive performance on SDXL, demonstrating its strong transferability.

C MORE VISUAL RESULTS

Additional visual demonstrations of our proposed ETCE method are presented in Figures 1 through 3.

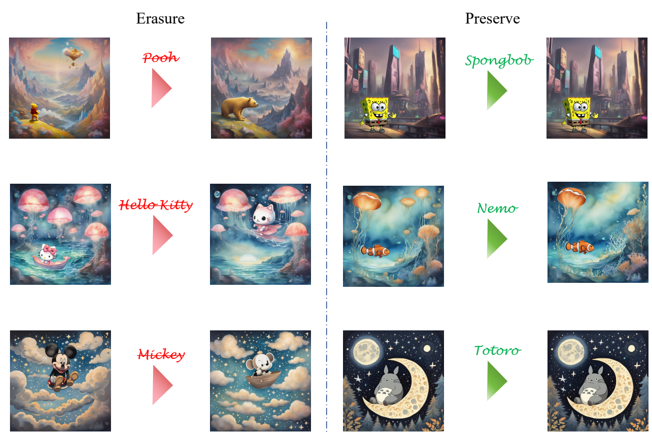

Figure 1: Instance erasure guided by fine-grained textual descriptions of artistic attributes. Our work achieves high-quality concept removal for multiple objects while maintaining minimal risk of unintended erasure to preserved non-target concepts.