

A BROADER IMPACTS AND LIMITATIONS

The abundance of training data not only enhances the performance of generative models but also introduces issues with privacy, unfairness, and bias. Our proposed controllable unlearning framework offers a viable solution to these issues. Our proposed framework is not limited to unlearning in I2I generation models but can be easily extended to other types of generative models, including text-to-image and text-to-text models. However, the unlearning framework presented herein has certain limitations. Note that Propositions 1 and 2 in Section 4 assume the convexity of the objective function and the feasible set. This assumption is essential to guarantee that the yielded solutions are Pareto optimal. In cases where the objective function and the feasible set are non-convex, the solutions obtained from solving Eq. (6) can only be guaranteed to be weakly Pareto optimal (Miettinen, 1999) or Pareto stability (Chen et al., 2024).

B DISCUSSION ON THE OBJECTIVE OF UNLEARNING

Describing the unlearning target as inpainting an image using only background content is feasible to some extent, such as concept unlearning. For instance, if we aim to protect privacy by unlearning parts of an image generation model that contain personal information (i.e., an abstract concept), we can first identify the region of the image containing such information, then simply mask this region, and subsequently generate a new image through inpainting, ensuring that the model’s output aligns with the inpainted new image. However, this approach has two issues:

- Firstly, it must be ensured that the new image generated through inpainting does not contain the information that needs to be forgotten. We believe this can be accomplished by incorporating an additional adversarial discriminator using GAN training strategies or by employing reinforcement strategies.
- Secondly, aligning the model’s output with the inpainted new image merely confuses the knowledge learned by the model, increasing uncertainty during generation, which constitutes a superficial form of unlearning. However, based on our experimental experience, if the goal is merely to erase the influence of certain samples on the model, directly aligning with Gaussian noise may yield a more pronounced unlearning effect.

C THEORETICAL VALIDATION

C.1 PROOF OF EQUIVALENCE

Given the original problem

$$\min_{\theta \in \mathbb{R}^d} f_2(\theta) \quad \text{s.t.} \quad f_1(\theta) \leq \varepsilon, \quad (10)$$

which is a constrained nonlinear programming problem. To solve it, we formulate its Lagrangian equation:

$$\mathcal{L}(\theta, \lambda) = f_2(\theta) + \lambda(f_1(\theta) - \varepsilon). \quad (11)$$

Further, we derive the KKT conditions for Eq.11:

$$\begin{aligned} \nabla_{\theta} \mathcal{L}(\theta^*, \lambda^*) &= \nabla f_2(\theta^*) + \lambda^* \nabla[f_1(\theta^*) - \varepsilon] = 0 \\ f_1(\theta^*) - \varepsilon &\leq 0 \\ \lambda^* &\geq 0 \\ \lambda^*(f_1(\theta^*) - \varepsilon) &= 0. \end{aligned} \quad (12)$$

The standard Newton’s Method searches for the solution $\mathcal{L}_{\theta}(\theta, \lambda) = 0$ by iterating the following equation:

$$\begin{bmatrix} \theta_{t+1} \\ \lambda_{t+1} \end{bmatrix} = \begin{bmatrix} \theta_t \\ \lambda_t \end{bmatrix} - \underbrace{\begin{bmatrix} \nabla_{\theta}^2 \mathcal{L} & \nabla[f_1(\theta_t) - \varepsilon] \\ \nabla[f_1(\theta_t) - \varepsilon]^T & 0 \end{bmatrix}^{-1}}_{\nabla^2 \mathcal{L}^{-1}} \underbrace{\begin{bmatrix} \nabla_{\theta} \mathcal{L}(\theta_t, \lambda_t) \\ f_1(\theta_t) - \varepsilon \end{bmatrix}}_{\nabla \mathcal{L}}, \quad (13)$$

where ∇_{θ}^2 denotes the Hessian matrix. However, the Newton step $g_t = (\nabla_{\theta}^2 \mathcal{L})^{-1} \nabla_{\theta} \mathcal{L}$ cannot be calculated directly and we also have other optimal condition in Eq. 12 introduced by the inequality constraints. Instead, the basic sequential quadratic programming algorithm defines an appropriate search direction g_t at an iterate (θ_t, λ_t) , as a solution to the quadratic programming subproblem.

Denoting by $g_t = (g_t^{\theta}, g_t^{\lambda})$ the change in the variables at the current point (θ_t, λ_t) , where $(g_t^{\theta}, g_t^{\lambda})$ solve the Newton-KKT system (Nocedal & Wright):

$$\begin{aligned} \nabla_{\theta} \mathcal{L}(\theta_t, \lambda_t) g_t^{\theta} + \nabla[f_1(\theta_t) - \varepsilon] g_t^{\lambda} &= -\nabla_{\theta} \mathcal{L}(\theta_t, \lambda_t) \\ f_1(\theta_t) - \varepsilon + \nabla[f_1(\theta_t) - \varepsilon] g_t^{\theta} &\leq 0 \\ \lambda_t + g_t^{\lambda} &\geq 0 \\ (\lambda_t + g_t^{\lambda}) (f_1(\theta^*) - \varepsilon + \nabla[f_1(\theta_t) - \varepsilon] g_t^{\theta}) &= 0. \end{aligned} \quad (14)$$

Denoting by $\lambda_{t+1} = \lambda_t + g_t^{\lambda}$, we have

$$\begin{aligned} \nabla_{\theta} \mathcal{L}(\theta_t, \lambda_t) g_t^{\theta} + \nabla[f_1(\theta_t) - \varepsilon] \lambda_{t+1} &= -\nabla f_2(\theta_t) \\ f_1(\theta_t) - \varepsilon + \nabla[f_1(\theta_t) - \varepsilon] g_t^{\theta} &\leq 0 \\ \lambda_{t+1} &\geq 0 \\ \lambda_{t+1} (f_1(\theta^*) - \varepsilon + \nabla[f_1(\theta_t) - \varepsilon] g_t^{\theta}) &= 0. \end{aligned} \quad (15)$$

It is easy to check that Eq. 15 is the optimality system of the following quadratic problem (QP)

$$\begin{aligned} \min_g \quad & f_2(\theta_t) + \nabla f_2(\theta_t)^{\top} g + \frac{1}{2} g^{\top} \nabla_{\theta}^2 \mathcal{L}(\theta_t, \lambda_t) g \\ & f_1(\theta_t) - \varepsilon + \nabla[f_1(\theta_t) - \varepsilon] g \leq 0. \end{aligned} \quad (16)$$

Setting $g_t^{\theta} = g$, the KKT conditions for Eq. 16 are consistent with the constraints specified in Eq. 15. Further, according to Theorem 1, the optimal solution for Eq. 16, when approaching the optimal solution of the original Problem (i.e., Eq. 10), satisfies the KKT conditions of Eq. 10. Considering that the models discussed in this paper are all deep neural networks, based on previous studies (Welling & Teh, 2011; Martens, 2016; Zhang et al., 2021; 2022; 2024), the initial guess Hessian matrix can be approximated as an identity matrix. Additionally, for consistency with the main text (i.e., $\theta_{t+1} \leftarrow \theta_t - \mu_t g_t$), setting $g = -g_t$ yields the following form:

$$\begin{aligned} \min_{g_t} \quad & \nabla f_2(\theta_t)^{\top} \nabla f_2(\theta_t) - 2 \nabla f_2(\theta_t)^{\top} g_t + g_t^{\top} g_t \\ & \nabla f_1(\theta_t) g_t \geq f_1(\theta_t) - \varepsilon. \end{aligned} \quad (17)$$

Theorem 1. *Theorem of Robinson (1974). Suppose that θ^* is a local solution of Eq. 10 at which the KKT conditions are satisfied for some λ^* . Suppose, too, that the linear independence constraint qualification (LICQ), the strict complementarity condition, and the second-order sufficient conditions hold at (θ^*, λ^*) . Then if (θ_t, λ_t) is sufficiently close to (θ^*, λ^*) , there is a local solution of the subproblem Eq. 16 whose active set \mathcal{A}_t is the same as the active set $\mathcal{A}(\theta^*)$ of the nonlinear program Eq. 10 at θ^* .*

C.2 BASIC COMPONENTS

Before exploring the proofs of Propositions 1 and 2, it is essential to define some fundamental concepts and lemmas. This references some works (Boyd & Vandenberghe, 2004; Pardalos et al., 2017; Gong et al., 2021) mentioned earlier; for the sake of readability, we will reiterate them here.

Penalty Function. An alternative method to evaluate the optimality of Algorithm 1 involves the L_1 penalty function given by:

$$P_{\xi}(\theta) = f_2(\theta) + \xi[f_1(\theta) - \varepsilon]_+, \quad (18)$$

where $\xi > 0$ is a scaling coefficient. The minima of Eq. (18) align with the solutions to Eq. (6) for sufficiently large values of ξ (Nocedal & Wright).

First-order KKT Condition and KKT Function. We revisit the first-order KKT condition (Nocedal & Wright) for the constrained optimization described in Eq. (9). Assume θ^* is a local optimum

with continuously differentiable $f_1(\theta)$ and $f_2(\theta)$, and $\|\nabla f_1(\theta^*)\| \neq 0$. There exists a Lagrange multiplier $\omega^* \in [0, +\infty)$ such that:

$$\nabla f_2(\theta^*) + \omega^* \nabla f_1(\theta^*) = 0, \quad f_1(\theta^*) \leq \varepsilon, \quad \omega^*(f_1(\theta^*) - \varepsilon) = 0. \quad (19)$$

This setup highlights the importance of $\|\nabla f_1(\theta^*)\| \neq 0$ as a constraint qualification condition.

Utilizing Algorithm 1 for Eq. (9), and for $\eta \geq 0$, the KKT function (Gong et al., 2021) to verify the first-order KKT condition is defined as:

$$K_\tau(\theta_t, \eta_t) = \|\nabla f_2(\theta_t) + \eta_t \nabla f_1(\theta_t)\|^2 + \tau[\psi(\theta_t)]_+ + \eta_t[-\psi(\theta_t)]_+, \quad (20)$$

where $\tau > 0$, and $[x]_+ = \max(x, 0)$. It is clear that $K_\tau(\theta_t, \eta_t) \geq 0$ for all $\theta_t \in \mathbb{R}^d$ and $\eta_t \geq 0$, achieving $K_\tau(\theta_t, \eta_t) = 0$ iff (θ_t, η_t) satisfies the first-order KKT condition.

Second-order KKT Condition and KKT Function

In the context of Algorithm 1 applied to Eq. (8), we expect that $\|\nabla f_1(\theta_t)\|$ approaches zero, leading to η_t potentially diverging to infinity. This scenario indicates a violation of the first-order KKT condition, potentially interpreted as $\eta^* = +\infty$.

While the first-order condition (Eq. (19)) is inadequate, the second-order KKT conditions involving the Hessian $\nabla^2 f_1(\theta)$ are applicable (Dempe et al., 2010). Consider the relaxed form of Eq. (8) as:

$$\min_{\theta \in \mathbb{R}^d} f_2(\theta) \quad \text{s.t.} \quad \nabla f_1(\theta) = 0. \quad (21)$$

If θ^* is a local minimum of Eq. (8), it coincides with a local minimum of Eq. (21). Assuming $f_2(\theta)$ and $\nabla f_1(\theta)$ are continuously differentiable, with the Hessian $\nabla^2 f_1(\theta)$ maintaining constant rank near θ^* (Janin, 1984), the first-order KKT condition for Eq. (21) can be formulated. There exists a vector $\omega^* \in \mathbb{R}^d$ such that:

$$\nabla f_2(\theta^*) + \nabla^2 f_1(\theta^*) \omega^* = 0. \quad (22)$$

This condition implies that $\nabla f_2(\theta^*)$ is orthogonal to the null space of $\nabla^2 f_1(\theta^*)$, defining the tangent space of the stationary manifold $\{\theta : \nabla f_1(\theta) = 0\}$ for $f_1(\theta)$.

For verifying local optimality under the constraints of Eq. (8) where $\psi(\theta) \geq 0$, the KKT function is proposed as:

$$K_\tau(\theta_t, \eta_t) = \|\nabla f_2(\theta_t) + \eta_t \nabla f_1(\theta_t)\|^2 + \tau \psi(\theta_t), \quad (23)$$

where $\psi(\theta_t) = 0$ asserts that θ_t is stationary for $f_1(\theta)$, and $\|\nabla f_2(\theta_t) + \eta_t \nabla f_1(\theta_t)\| = 0$ signifies local optimality with respect to $f_2(\theta)$, aligning with the KKT condition for the relaxed problem $\min_{\theta} \{f_2(\theta) \text{ s.t. } f_1(\theta) \leq \varepsilon_t\}$, with $\varepsilon_t = f_1(\theta_t)$.

In the analysis of Algorithm 1, a fundamental theorem concerning the behavior of the penalty function $P_\xi(\theta)$ and the KKT function $K_\tau(\theta, \eta)$, given in Eqs. (20) and (23), is essential for understanding the algorithm's convergence and feasibility characteristics. This lemma is stated as follows:

Theorem 2. *Theorem 3.2 of Gong et al. (2021). Assume Assumption 1 holds, for any $\xi \geq 0$, we have*

$$\frac{d}{dt} P_\xi(\theta_t) \leq -K_{\xi-\eta_t}(\theta_t, \eta_t), \forall t \in [0, +\infty). \quad (24)$$

This equation indicates that $P_\xi(\theta_t)$ is non-increasing w.r.t. time t provided that $K_{\xi-\eta_t}(\theta_t, \eta_t) \geq 0$. This condition is satisfied if ξ is sufficiently large such that $\xi - \eta_t \geq 0$, or when the constraint is met, i.e., $f_1(\theta_t) \leq \varepsilon$, ensuring $[\psi(\theta_t)]_+ = 0$.

This lemma facilitates further deductions about the behavior of the algorithm under different settings of the parameter ξ . For instance, setting $\xi \rightarrow +\infty$ allows us to demonstrate that the constraint $[f_1(\theta_t) - \varepsilon]_+$ is non-increasing w.r.t. time t . This implies that $f_1(\theta_t)$ is decreasing w.r.t. time t outside the feasible region, and once θ_t enters the feasible region, it remains therein. Conversely, setting $\xi = 0$ reveals that $f_2(\theta_t)$ monotonically decreases w.r.t. time t within the feasible set, progressing towards a KKT point. These observations are critical for understanding both the feasibility and optimality properties of Algorithm 1 under different operational scenarios.

Proposition 4. *Under Assumption 1, the following two propositions hold:*

1. *For any time $t \in [0, +\infty)$, $\min_{s \in [0, t]} [\psi(\theta_s)]_+ = O(\frac{1}{t})$ holds.*

2. If $\psi(\theta) \geq 0$ holds, then $\min_{s \in [0, t]} \psi(\theta_s) \leq \frac{1}{t} (f_1(\theta_0) - f_1^*)$ for any time $t \in [0, +\infty)$.

Proof of Proposition 4-1. At each time point $t \in [0, +\infty)$, dividing both sides of Eq. (24) by $\xi > 0$ and taking $\xi \rightarrow +\infty$ gives

$$\frac{d}{dt} [f_1(\theta_t) - \varepsilon]_+ \leq -[\psi(\theta_t)]_+ \leq 0.$$

Integrating this on time interval $[0, t]$ gives

$$\begin{aligned} \min_{s \in [0, t]} [\psi(\theta_s)]_+ &\leq \frac{1}{t} \int_0^t [\psi(\theta_s)]_+ ds \\ &\leq \frac{1}{t} ([f_1(\theta_0) - \varepsilon]_+ - [f_1(\theta_t) - \varepsilon]_+) \\ &\leq \frac{1}{t} [f_1(\theta_0) - \varepsilon]_+. \end{aligned} \quad (25)$$

$$\min_{s \in [0, t]} [\psi(\theta_s)]_+ \leq \frac{1}{t} \int_0^t [\psi(\theta_s)]_+ ds \leq \frac{1}{t} ([f_1(\theta_0) - \varepsilon]_+ - [f_1(\theta_t) - \varepsilon]_+) \leq \frac{1}{t} [f_1(\theta_0) - \varepsilon]_+.$$

This implies that $\min_{s \in [0, t]} [\psi(\theta_s)]_+ = O(\frac{1}{t})$. \square

Proof of Proposition 4-2. Let $f_1^* = \inf_{\theta \in \mathbb{R}^d} f_1(\theta)$ and $f_2^* = \inf_{\theta \in \mathbb{R}^d} f_2(\theta)$. Since $\psi(\theta) \geq 0$, by substituting Eq. (23) into Eq. (24), we have for any $\xi \geq 0$,

$$\frac{d}{dt} (f_2(\theta_t) + \xi [f_1(\theta_t) - \varepsilon]_+) \leq -\|\nabla f_2(\theta_t) + \eta_t \nabla f_1(\theta_t)\|^2 - (\xi - \eta_t) \psi(\theta_t), \quad \forall t \in [0, +\infty).$$

Integrating both sides from 0 to t yields:

$$\begin{aligned} \int_0^t (\|\nabla f_2(\theta_s) + \eta_s \nabla f_1(\theta_s)\|^2 + (\xi - \eta_s) \psi(\theta_s)) ds &\leq (f_2(\theta_0) - f_2(\theta_t)) + \xi [f_1(\theta_0) - \varepsilon]_+ - \xi [f_1(\theta_t) - \varepsilon]_+ \\ &\leq (f_2(\theta_0) - f_2^*) + \xi (f_1(\theta_0) - f_1^*). \end{aligned} \quad (26)$$

Taking $\xi \rightarrow +\infty$ in Eq. (26) gives

$$\int_0^t \psi(\theta_s) ds \leq f_1(\theta_0) - f_1^*, \quad (27)$$

which implies $\min_{s \in [0, t]} \psi(\theta_s) \leq \frac{1}{t} \int_0^t \psi(\theta_s) ds \leq \frac{1}{t} (f_1(\theta_0) - f_1^*)$. \square

C.3 PROOF OF PROPOSITION 1

Proof of Proposition 1. As θ_t converges to θ^* for $t \rightarrow +\infty$ and given the continuity of $\psi(\theta)$ and $\nabla f_1(\theta)$, it follows that $\lim_{t \rightarrow +\infty} \psi(\theta_t) = \psi(\theta^*)$, and $\lim_{t \rightarrow +\infty} \|\nabla f_1(\theta_t)\| = \|\nabla f_1(\theta^*)\|$.

Given $\psi(\theta) \geq 0$ and $\varepsilon = f_1^*$, Eq. (26) establishes that $\int_0^{+\infty} \psi(\theta_t) dt \leq f_1(\theta_0) - f_1^* < +\infty$. Consequently, $\lim_{t \rightarrow +\infty} \psi(\theta_t) = \psi(\theta^*) = 0$.

Given θ^* as a limit point of $\{\theta_t\}$, there exists an increasing sequence $\{t_n : n = 1, 2, \dots\}$ such that $t_n \rightarrow +\infty$ and $\theta_{t_n} \rightarrow \theta^*$ as $n \rightarrow +\infty$. The continuity of $\psi(\theta)$ and $\nabla f_1(\theta)$ ensures $\lim_{n \rightarrow +\infty} \psi(\theta_{t_n}) = \psi(\theta^*) = 0$, and $\lim_{n \rightarrow +\infty} \|\nabla f_1(\theta_{t_n})\| = \|\nabla f_1(\theta^*)\|$.

Since $\psi(\theta^*) = 0$ and the sign condition of $\psi(\theta)$, it implies $\text{sign}(f_1(\theta^*) - f_1^*) = \text{sign}(\psi(\theta^*)) = 0$. Therefore $f_1(\theta^*) = f_1^*$ and θ^* is a minimum point of $f_1(\theta)$. This gives $\lim_{n \rightarrow +\infty} \|\nabla f_1(\theta_{t_n})\| = \|\nabla f_1(\theta^*)\| = 0$.

Given $\lim_{t \rightarrow +\infty} g_t = 0$, we deduce that $\lim_{t \rightarrow +\infty} \|\nabla f_2(\theta_t) + \eta_t \nabla f_1(\theta_t)\| = \lim_{t \rightarrow +\infty} \|g_t\| = 0$. Additionally, employing $\psi(\theta) \geq 0$, Eq. (23) implies $\lim_{t \rightarrow +\infty} K_\tau(\theta_t, \eta_t) = 0$ for some $\tau > 0$.

Combining $\lim_{t \rightarrow +\infty} \|\nabla f_2(\theta_t) + \eta_t \nabla f_1(\theta_t)\| = 0$ and $\nabla f_1(\theta^*) = \lim_{n \rightarrow +\infty} \nabla f_1(\theta_{t_n}) = 0$, we can derive

$$\begin{aligned} \|\nabla f_2(\theta_t) + \eta_t \nabla f_1(\theta_t)\| &= \|\nabla f_2(\theta_t) + \eta_t (\nabla f_1(\theta_t) - \nabla f_1(\theta^*))\| \\ &= \|\nabla f_2(\theta_t) + \eta_t \nabla^2 f_1(\theta'_t) (\theta_t - \theta^*)\| \\ &= \|\nabla f_2(\theta_t) + \nabla^2 f_1(\theta'_t) \omega'_t\|. \end{aligned}$$

where θ'_t is a convex combination of θ_t and θ^* , and we defined $\omega'_t = \eta_t (\theta_t - \theta^*)$.

Define $\omega_t = (\nabla^2 f_1(\theta'_t))^+ \nabla f_2(\theta_t)$, where $(\nabla^2 f_1(\theta'_t))^+$ denotes the Moore-Penrose pseudo-inverse of matrix $\nabla^2 f_1(\theta'_t)$, which satisfies that

$$\omega_t = \arg \min_{\omega \in \mathbb{R}^d} \left\{ \|\omega\| \quad \text{s.t.} \quad \omega \in \arg \min_w \|\nabla f_2(\theta_t) + \nabla^2 f_1(\theta'_t) \omega\| \right\}.$$

It follows that

$$\|\nabla f_2(\theta_t) + \nabla^2 f_1(\theta'_t) \omega_t\| \leq \|\nabla f_2(\theta_t) + \nabla^2 f_1(\theta_t) \omega'_t\| = \|\nabla f_2(\theta_t) + \eta_t \nabla f_1(\theta_t)\|.$$

Given $\|\nabla f_2(\theta_{t_n}) + \eta_{t_n} \nabla f_1(\theta_{t_n})\| \rightarrow 0$ as $n \rightarrow +\infty$, we have $\|\nabla f_2(\theta_{t_n}) + \nabla^2 f_1(\theta'_{t_n}) \omega_{t_n}\| \rightarrow 0$. Assuming $\theta_{t_n} \rightarrow \theta^*$ and $\theta'_{t_n} \rightarrow \theta^*$ as $n \rightarrow +\infty$, and by the constant rank condition and relevant corollary of Stewart (1977) (rephrased in Lemma 2), we deduce $(\nabla^2 f_1(\theta'_{t_n}))^+ \rightarrow (\nabla^2 f_1(\theta^*))^+$ and hence $\omega_{t_n} \rightarrow \omega^*$ as $n \rightarrow +\infty$, where $\omega^* := (\nabla^2 f_1(\theta^*))^+ \nabla f_2(\theta^*)$. Thus, $\|\nabla f_2(\theta_t) + \nabla^2 f_1(\theta'_t) \omega_t\| \rightarrow \|\nabla f_2(\theta^*) + \nabla^2 f_1(\theta^*) \omega^*\|$, leading to $\|\nabla f_2(\theta^*) + \nabla^2 f_1(\theta^*) \omega^*\| = 0$, which implies that θ^* satisfies the second-order KKT conditions for Eq. (22).

Given the convexity of $f_1(\theta)$ and $f_2(\theta)$ with respect to θ , then $f_2(\theta^*)$ is the minimum in the feasible set $\Omega = \{\theta : f_1(\theta) \leq \varepsilon\}$, without any $\hat{\theta} \in \Omega$ such that $f_2(\hat{\theta}) < f_2(\theta^*)$. Consequently, θ^* is a solution to Eq. (8). According to Chankong and Haimes (Chankong & Haimes, 1982), this solution is unique without further checking, as affirmed by theorem of Miettinen (rephrased in Lemma 3), θ^* is Pareto optimal.

Therefore, combining the conclusions, θ^* is established as both the minimum of $f_1(\theta)$ and Pareto optimal, confirming its status as Pareto optimal for complete unlearning.

□

Lemma 2. *Corollary 3.5 of Stewart (1977). Let $\{A_t\}$ be a sequence of matrices converging to A_* as $t \rightarrow +\infty$. The condition $\lim_{t \rightarrow +\infty} A_t^+ = A_*^+$ is equivalent to the condition that $\text{rank}(A_t) = \text{rank}(A_*)$ for all t sufficiently large.*

Lemma 3. *Theorem 3.2.4 of Miettinen (1999). A point $\theta^* \in \Omega$ is Pareto optimal if it is a unique solution of ε -constraint problem (Eq. (6)) for any given upper bound vector $\varepsilon = (\varepsilon_1, \dots, \varepsilon_{\ell-1}, \varepsilon_{\ell+1}, \dots, \varepsilon_t)^T$.*

C.4 PROOF OF PROPOSITION 2

Proof of Proposition 2. Since θ_t is stationary, $\dot{\theta}_t = -g_t = 0$, implying $\frac{d}{dt} P_\xi(\theta_t) = 0$ for all $\xi \geq 0$. From Eq. (24), we have $\frac{d}{dt} P_\xi(\theta_t) \leq -K_{\xi-\eta_t}(\theta_t, \eta_t)$. Consequently, $K_{\xi-\eta_t}(\theta_t, \eta_t) \leq 0$ for all $\xi \geq \eta_t$. Setting $\xi = \eta_t + \tau$, where $\tau \geq 0$, it follows that $K_\tau(\theta_t, \eta_t) = 0$. This implies that θ^* satisfies the first-order KKT conditions for Eq. (19), i.e., there exists a Lagrange multiplier $\eta^* \in [0, +\infty)$ such that

$$\nabla f_2(\theta^*) + \eta^* \nabla f_1(\theta^*) = 0, \quad f_1(\theta^*) \leq \varepsilon, \quad \eta^* (f_1(\theta^*) - \varepsilon) = 0.$$

As affirmed by theorem of Miettinen (rephrased in Lemma 4), θ_t is a Pareto optimal solution.

□

Lemma 4. *Theorem 3.1.8 of Miettinen (1999). (Karush-Kuhn-Tucker sufficient condition for Pareto optimality) Let the objective and the constraint functions of problem Eq. (9) be convex and continuously differentiable at a decision vector $\theta^* \in \Omega$. A sufficient condition for θ^* to be Pareto optimal is that there exist multipliers $\mu^* > \mathbf{0}$ and $\eta^* > \mathbf{0}$ such that*

$$\begin{aligned} (1) \quad & \mu^* \nabla f_2(\theta^*) + \eta^* \nabla f_1(\theta^*) = \mathbf{0} \\ (2) \quad & \eta^* (f_1(\theta^*) - \varepsilon) = 0. \end{aligned}$$

C.5 PROOF OF PROPOSITION 3

Proof of Proposition 3-1. Given that $\psi(\theta) \geq 0$, we recall conclusions from Proposition 4-2:

$$\min_{s \in [0, t]} \psi(\theta_s) \leq \frac{1}{t} (f_1(\theta_0) - f_1^*).$$

Taking $\xi = 0$ in Eq. (26) gives

$$\int_0^t \|\nabla f_2(\theta_s) + \eta_s \nabla f_1(\theta_s)\|^2 ds \leq \int_0^t \eta_s \psi(\theta_s) ds + (f_2(\theta_0) - f_2^*).$$

To derive an upper bound for $\int_0^t \|\nabla f_2(\theta_s) + \eta_s \nabla f_1(\theta_s)\|^2 ds$, the principal challenge lies in bounding $\int_0^t \eta_s \psi(\theta_s) ds$.

Given the assumption $0 \leq \psi(\theta_t) \leq \alpha \|\nabla f_1(\theta_t)\|^\delta$, where $\delta \geq 1$, and applying Lemma 5, we obtain:

$$\eta_t \psi(\theta_t) \leq \left(\alpha \|\nabla f_1(\theta_t)\|^{\delta-1} + \|\nabla f_2(\theta_t)\| \right) \alpha^{\frac{1}{\delta}} \psi(\theta_t)^{1-\frac{1}{\delta}} \leq \Upsilon \psi(\theta_t)^{1-\frac{1}{\delta}},$$

where $\Upsilon = \sup_{\theta \in \mathbb{R}^d} \left(\alpha \|\nabla f_1(\theta)\|^{\delta-1} + \|\nabla f_2(\theta)\| \right) \alpha^{\frac{1}{\delta}}$. This leads to

$$\begin{aligned} \int_0^t \eta_s \psi(\theta_s) ds &\leq \Upsilon \int_0^t \psi(\theta_s)^{1-\frac{1}{\delta}} ds \\ &\leq \Upsilon \left(\int_0^t \psi(\theta_s) ds \right)^{1-\frac{1}{\delta}} \left(\int_0^t 1 ds \right)^{\frac{1}{\delta}} \\ &\leq \Upsilon \left(\int_0^t \psi(\theta_s) ds \right)^{1-\frac{1}{\delta}} t^{\frac{1}{\delta}} \\ &\leq \Upsilon (f_1(\theta_0) - f_1^*)^{1-\frac{1}{\delta}} t^{\frac{1}{\delta}}. \end{aligned}$$

Thus, it follows that

$$\begin{aligned} \min_{s \in [0, t]} \|\nabla f_2(\theta_s) + \eta_t \nabla f_1(\theta_s)\|^2 &\leq \frac{1}{t} \int_0^t \|\nabla f_2(\theta_s) + \eta_t \nabla f_1(\theta_s)\|^2 ds \\ &\leq \frac{1}{t} \int_0^t \eta_s \psi(\theta_s) dt + \frac{1}{t} (f_2(\theta_0) - f_2^*) \\ &\leq \frac{1}{t^{1-\frac{1}{\delta}}} \Upsilon (f_1(\theta_0) - f_1^*)^{1-\frac{1}{\delta}} + \frac{1}{t} (f_2(\theta_0) - f_2^*). \end{aligned}$$

Since $\|g_s\|^2 = \|\nabla f_2(\theta_s) + \eta_t \nabla f_1(\theta_s)\|^2$, for any time $t \in [0, +\infty)$, we derive the following inequalities

$$\min_{s \in [0, t]} \|g_s\|^2 \leq \Upsilon \left(\frac{f_1(\theta_0) - f_1^*}{t} \right)^{1-\frac{1}{\delta}} + \frac{f_2(\theta_0) - f_2^*}{t}.$$

Combine the conclusions above, if $\psi(\theta) = \alpha \|\nabla f_1(\theta)\|^\delta$, we can further assert $\min_{s \in [0, t]} \|\nabla f_1(\theta_s)\| = O\left(1/t^{\frac{1}{\delta}}\right)$ and $\min_{s \in [0, t]} \|g_s\| = O\left(1/t^{\frac{1}{2}-\frac{1}{2\delta}}\right)$. Hence, the exponent δ controls the convergence rates of $\|\nabla f_1(\theta_t)\|$ (measuring the minimization of $f_1(\theta)$), and that of $\|g_t\|$ (measuring the minimization of $f_2(\theta)$).

If $\psi(\theta) = c \ln(1 + \gamma \|\nabla f_1(\theta)\|)$, where $c\gamma \leq \alpha$ and $\delta = 1$. Since $0 \leq \psi(\theta) \leq \alpha \|\nabla f_1(\theta)\|^\delta$, the assumptions of Proposition 4 are satisfied. Consequently, we obtain $\min_{s \in [0, t]} \|\nabla f_1(\theta_s)\| = O\left(e^{\frac{1}{ct}}\right)$ and $\min_{s \in [0, t]} \|g_s\|^2 = O\left(1/t^{\frac{1}{2} - \frac{1}{2\delta}}\right)$.

□

Proof of Proposition 3-2. If $\psi(\theta) = \beta(f_1(\theta) - \varepsilon)^\delta$, where $\delta = 2n + 1$, $n \in \mathbb{N}$. According to Proposition 4-1, we deduce that $\min_{s \in [0, t]} \beta[(f_1(\theta_s) - \varepsilon)^\delta]_+ \leq \frac{1}{t} [f_1(\theta_0) - \varepsilon]_+$. Consequently, we obtain $\min_{s \in [0, t]} [f_1(\theta_s) - \varepsilon]_+ = O\left(1/t^{\frac{1}{\delta}}\right)$.

□

Lemma 5. Rewrite based on Lemma 6.1 of Gong et al. (2021). Let $\eta_t = \min_{\eta \geq 0} \left\{ \|\nabla f_2(\theta_t) + \eta \nabla f_1(\theta_t)\|^2 - \eta \psi(\theta_t) \right\} = \max\left(\frac{\psi(\theta_t) - \nabla f_2(\theta_t)^\top \nabla f_1(\theta_t)}{\|\nabla f_1(\theta_t)\|^2}, 0\right)$ and assume $0 \leq \psi(\theta_t) \leq \alpha \|\nabla f_1(\theta_t)\|^\delta$ for $\alpha \geq 0$ and $\delta \geq 1$. Then

$$\eta_t \psi(\theta_t) \leq (\alpha \|\nabla f_1(\theta_t)\|^\delta + \|\nabla f_2(\theta_t)\|) \alpha^{\frac{1}{\delta}} \psi(\theta_t)^{1 - \frac{1}{\delta}}. \quad (28)$$

Proof of Lemma 5. Given $\psi(\theta_t) \leq \alpha \|\nabla f_1(\theta_t)\|^\delta$, we have

$$\frac{\psi(\theta_t)}{\|\nabla f_1(\theta_t)\|} \leq \alpha \|\nabla f_1(\theta_t)\|^{\delta-1}, \text{ and } \frac{\psi(\theta_t)}{\|\nabla f_1(\theta_t)\|} \leq \alpha^{\frac{1}{\delta}} \psi(\theta_t)^{1 - \frac{1}{\delta}}.$$

With $\psi(\theta_t) \geq 0$, the upper bound for η_t simplifies to

$$\eta_t = \frac{\max(\psi(\theta_t) - \nabla f_2(\theta_t)^\top \nabla f_1(\theta_t), 0)}{\|\nabla f_1(\theta_t)\|^2} \leq \frac{\psi(\theta_t) + \|\nabla f_2(\theta_t)\| \|\nabla f_1(\theta_t)\|}{\|\nabla f_1(\theta_t)\|^2}.$$

Therefore,

$$\begin{aligned} \eta_t \psi(\theta_t) &\leq \frac{\psi(\theta_t)^2}{\|\nabla f_1(\theta_t)\|^2} + \|\nabla f_2(\theta_t)\| \frac{\psi(\theta_t)}{\|\nabla f_1(\theta_t)\|} \\ &\leq (\alpha \|\nabla f_1(\theta_t)\|^{\delta-1} + \|\nabla f_2(\theta_t)\|) \frac{\psi(\theta_t)}{\|\nabla f_1(\theta_t)\|} \\ &\leq (\alpha \|\nabla f_1(\theta_t)\|^{\delta-1} + \|\nabla f_2(\theta_t)\|) \alpha^{\frac{1}{\delta}} \psi(\theta_t)^{1 - \frac{1}{\delta}}. \end{aligned}$$

□

D MORE DETAILS OF EXPERIMENTS

D.1 HYPER-PARAMETER OF EXPERIMENTS

MAE. We set the learning rate to 10^{-4} with no weight decay. Both baselines and our method employ AdamW as the foundational optimizer with $\beta = (0.90, 0.95)$, with the distinction being that our method necessitates some improvements on the basic optimizer. We set the input image resolution to 224×224 and batch size to 32. Simultaneously, we set the coefficient of $\psi(\theta)$ in Phase I to $\alpha = 5$, and the coefficient of ψ in Phase II to $\beta = 5$, followed by training for 8 epochs. Overall, it takes an hour on an NVIDIA A40 (48G) server.

VQ-GAN. We set the learning rate to 10^{-4} with no weight decay. Both baselines and our method employ AdamW as the foundational optimizer with $\beta = (0.90, 0.95)$. Our method necessitates some improvements on the basic optimizer. We set the input image resolution to 256×256 and batch size to 16. Simultaneously, we set the coefficient of $\psi(\theta)$ in Phase I to $\alpha = 10$, and the coefficient of $\psi(\theta)$ in Phase II to $\beta = 10$, followed by training for 10 epochs. Overall, it takes two hours on an NVIDIA A40 (48G) server.

Diffusion model. We set the learning rate to 10^{-5} with no weight decay. Both baselines and our method employ Adam as the foundational optimizer. Our method necessitates some improvements on the basic optimizer. We set the input image resolution to 256×256 and batch size to 16. Simultaneously, we set the coefficient of $\psi(\theta)$ in Phase I to $\alpha = 1$, and the coefficient of $\psi(\theta)$ in Phase II to $\beta = 1$, followed by training for 4 epochs. Overall, it takes twelve hours on an NVIDIA A40 (48G) server.

D.2 EVALUATION METRICS

IS. Following (Li et al., 2024a), for ImageNet-1K, we directly use the Inception-v3 model checkpoint to calculate the IS score. For Places-365, we use the Resnet-50 model checkpoint to calculate IS scores (Zhou et al., 2017).

FID. Regardless of whether it is ImageNet-1K or Places-365, we directly use the Inception-v3 model checkpoint to calculate the FID score.

CLIP. Following (Li et al., 2024a), whether it is for ImageNet-1K or Places-365, we use the ViT-H-14 model checkpoint to calculate the clip embedding vectors of the generated images and the ground truth images (Radford et al., 2021). Afterward, we calculate the cosine similarity between the two vectors as the clip score.

E ROBUSTNESS TO RETAIN SAMPLES AVAILABILITY

In machine unlearning, sometimes the real retain samples are not available due to data retention policies. To tackle this challenge, following (Li et al., 2024a), we assess our method using images from other classes as substitutes for real retain samples. For instance, on ImageNet-1K, since we have already selected 200 classes, we randomly chose some images from the remaining 800 classes to act as a "proxy retain set" during the unlearning process. We incrementally reduce the proportion of real retain samples in the retain set and increased the proportion of proxy retain samples, with the experimental results presented in Table 3. As demonstrated, our method is largely unaffected by the reduced availability of retain samples, indicating robust performance.

Table 3: Results of center cropping 50% of the images under different retain set usage proportions. \uparrow indicates higher is better, and \downarrow indicates lower is better. 'F' and 'R' stand for the forget set and retain set, respectively. Here, all results are based on the solution with the highest degree of unlearning completeness in Phase I.

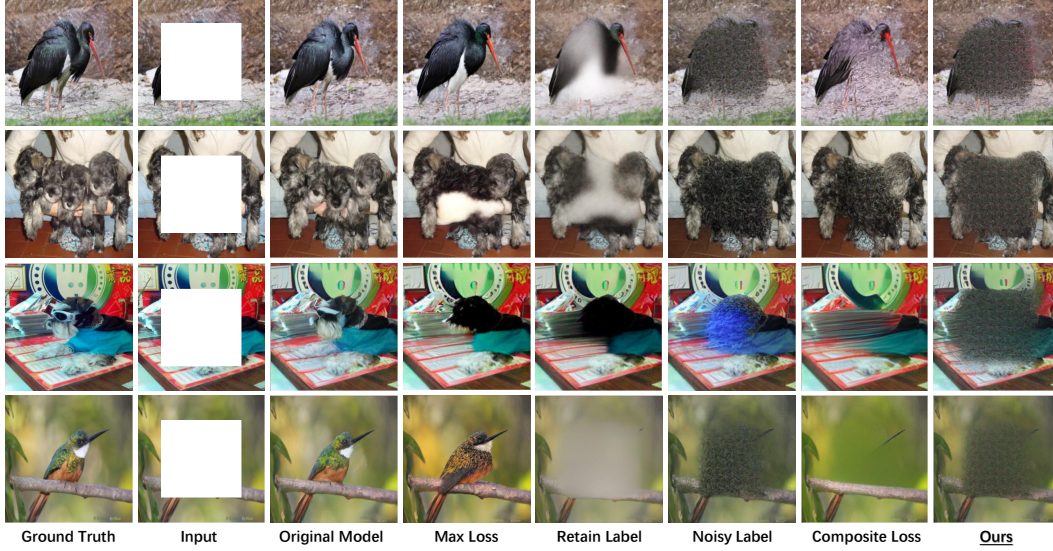
	MAE						VQ-GAN						Diffusion Models					
	IS			FID			IS			FID			IS			FID		
	F \downarrow	R \uparrow	F \downarrow	R \downarrow	F \downarrow	R \uparrow	F \downarrow	R \uparrow	F \downarrow	R \downarrow	F \downarrow	R \uparrow	F \downarrow	R \uparrow	F \downarrow	R \downarrow	F \downarrow	R \uparrow
Original	21.59	21.83	16.28	14.87	0.88	0.88	23.74	24.06	21.80	18.17	0.78	0.85	16.90	19.65	82.12	81.51	0.89	0.91
100%	12.33	17.47	154.60	68.453	0.69	0.75	13.23	22.55	139.21	26.39	0.46	0.82	11.84	18.47	165.05	95.42	0.55	0.81
80%	12.32	17.46	150.05	73.14	0.70	0.73	13.27	22.30	138.49	24.83	0.46	0.81	11.91	18.10	167.32	98.82	0.55	0.80
60%	12.22	17.42	150.55	74.22	0.70	0.73	13.24	22.54	140.35	24.92	0.61	0.81	12.06	18.53	165.24	98.43	0.60	0.80
40%	112.29	17.43	150.27	73.63	0.70	0.74	12.77	22.39	141.67	25.84	0.61	0.81	12.05	18.64	168.83	96.42	0.60	0.79
20%	12.50	17.68	147.45	70.75	0.70	0.74	12.77	22.39	144.38	28.08	0.60	0.81	13.49	18.67	168.26	95.47	0.57	0.79
0	12.21	17.68	147.31	68.09	0.70	0.74	12.39	22.35	147.17	29.79	0.62	0.80	13.24	18.76	168.43	96.63	0.60	0.79

F MORE GENERATED IMAGES: BASELINES VS OURS

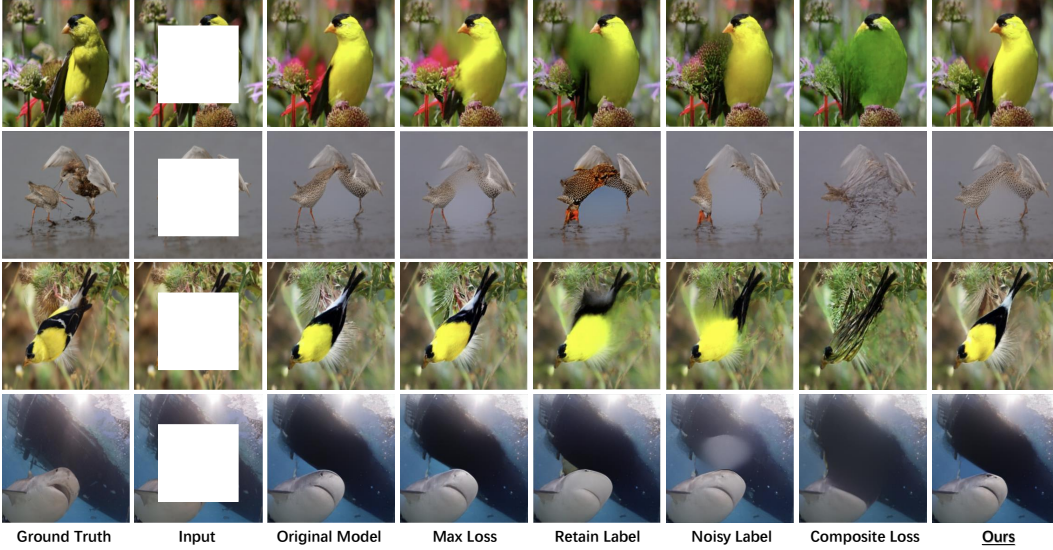
We conduct various generative tasks on three mainstream I2I generative models (i.e., MAE, VQ-GAN, and the diffusion model), including image expansion, inpainting, and reconstruction, to assess both baselines and our proposed method. Specifically, we conduct evaluations of image inpainting and expansion tasks on VQ-GAN, image reconstruction tasks on MAE, and image inpainting tasks on the diffusion model. The results indicate that our method can adapt to mainstream I2I generative models and various image generation tasks.

VQ-GAN. We conduct experiments on image inpainting and expansion task unlearning on VQ-GAN, where examples of the image inpainting tasks are illustrated in Figure 5, and examples of

image expansion can be referred to in Appendix H. Our unlearning method is effective for both image inpainting and image expansion tasks, and it significantly surpasses baselines.



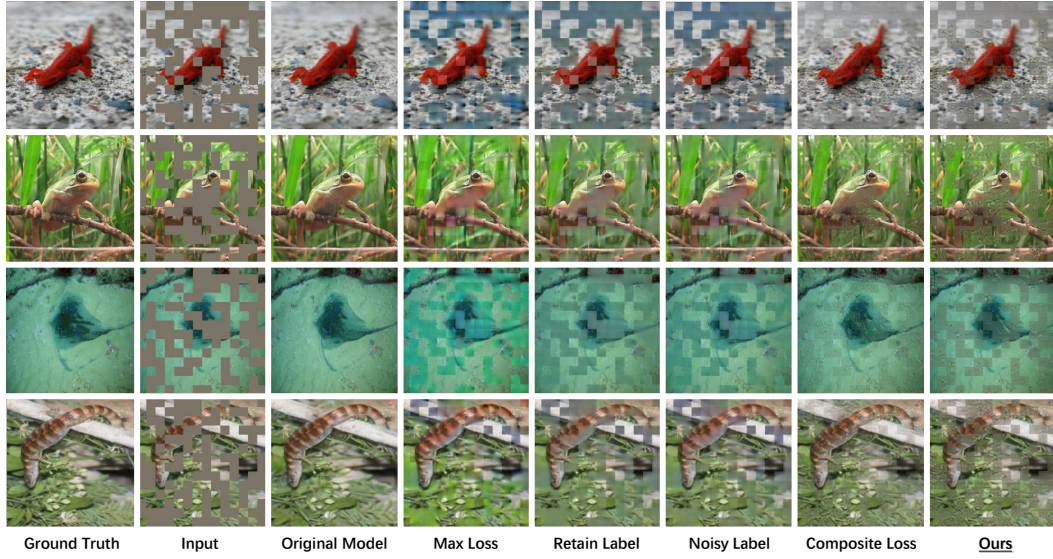
(a) Forget Set



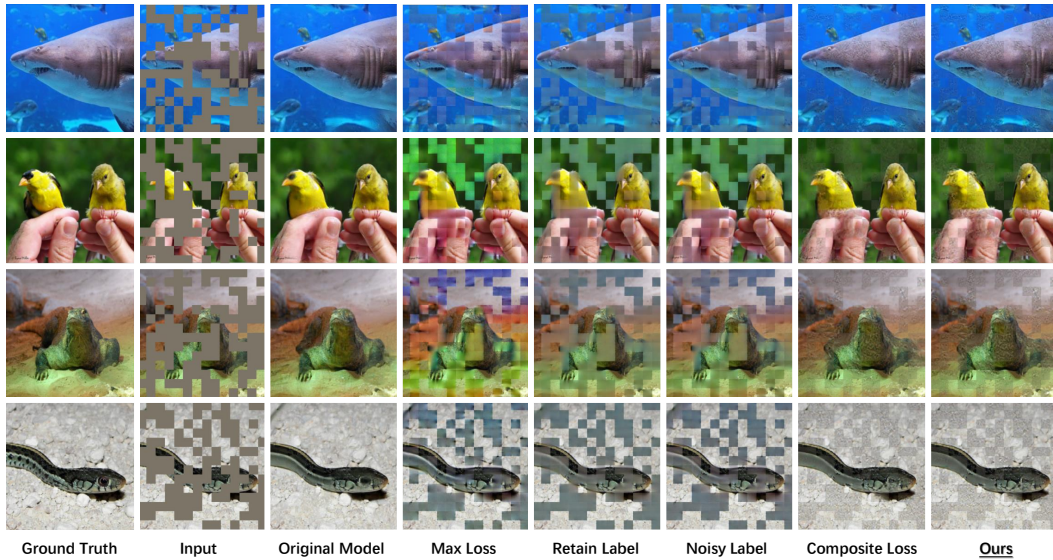
(b) Retain Set

Figure 5: VQ-GAN: generated images of cropping 50% at the center of the image. The upper part (a) represents the forget set, while the lower part (b) represents the retain set. "Ours" denotes the boundary condition of unlearning obtained in Phase I, which represents the point of the highest degree of unlearning completeness. It is evident that our method significantly outperforms baselines in terms of the unlearning effect on the forget set, most closely approximating Gaussian noise, and exhibits the least performance degradation on the retain set.

MAE. We conduct experiments on unlearning image reconstruction tasks on the MAE. As shown in figure 6, our unlearning method is also effective in the task of image reconstruction, with the effects of unlearning showing a significant advantage over baselines.



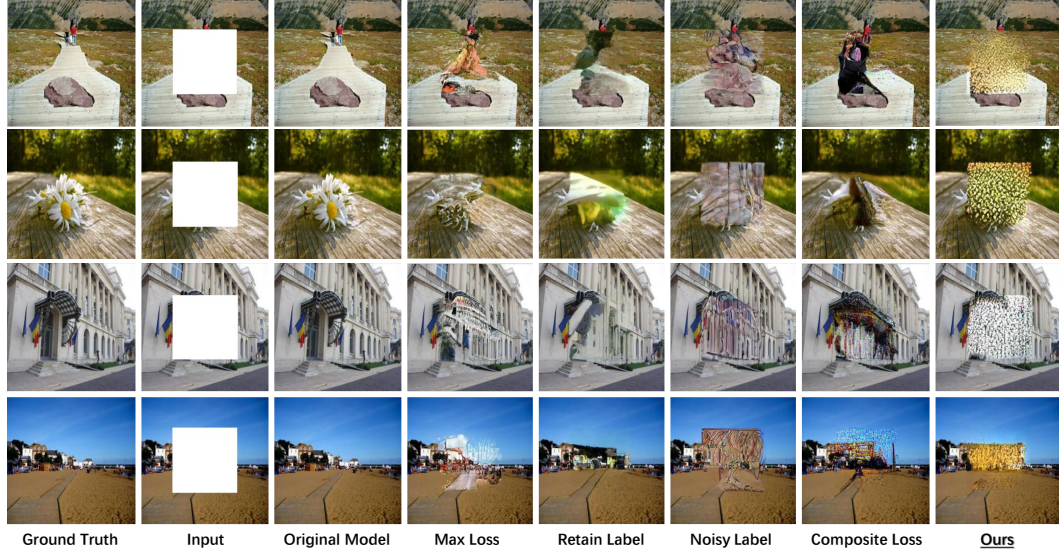
(a) Forget Set



(b) Retain Set

Figure 6: MAE: reconstruction of random masked images. We set the proportion of the random mask to 50%. The upper part (a) represents the forget set, while the lower part (b) represents the retain set. "Ours" denotes the boundary condition of unlearning obtained in Phase I, which represents the point of the highest degree of unlearning completeness.

Diffusion model. We validate our unlearning framework on the diffusion model task for image inpainting. As shown in figure 7, the results indicate that our method is equally applicable to diffusion models, and the effectiveness of unlearning surpasses that of baselines.



(a) Forget Set



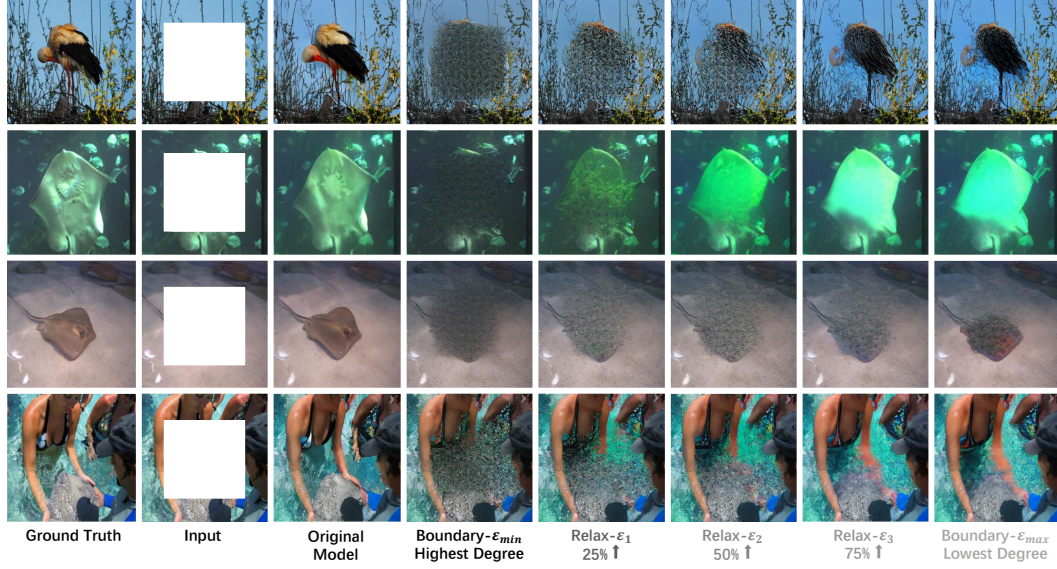
(b) Retain Set

Figure 7: Diffusion model: generated images of cropping 50% at the center of the image. The upper part (a) represents the forget set, while the lower part (b) represents the retain set. "Ours" denotes the boundary condition of unlearning obtained in Phase I, which represents the point of the highest degree of unlearning completeness.

G MORE GENERATED IMAGES: DIFFERENT DEGREES OF COMPLETENESS

We validate the control effect of our controllable unlearning framework across multiple generative tasks in three mainstream I2I generative models. The results demonstrate that our controllable unlearning framework can effectively control unlearning across various image generation tasks of mainstream I2I generative models.

VQ-GAN. We center-crop the image by 50% and utilize the VQ-GAN for image inpainting. Subsequently, we applied our unlearning framework to enforce unlearning. The results in Figure 8 demonstrate the effectiveness of our method, with the control effect being very pronounced.



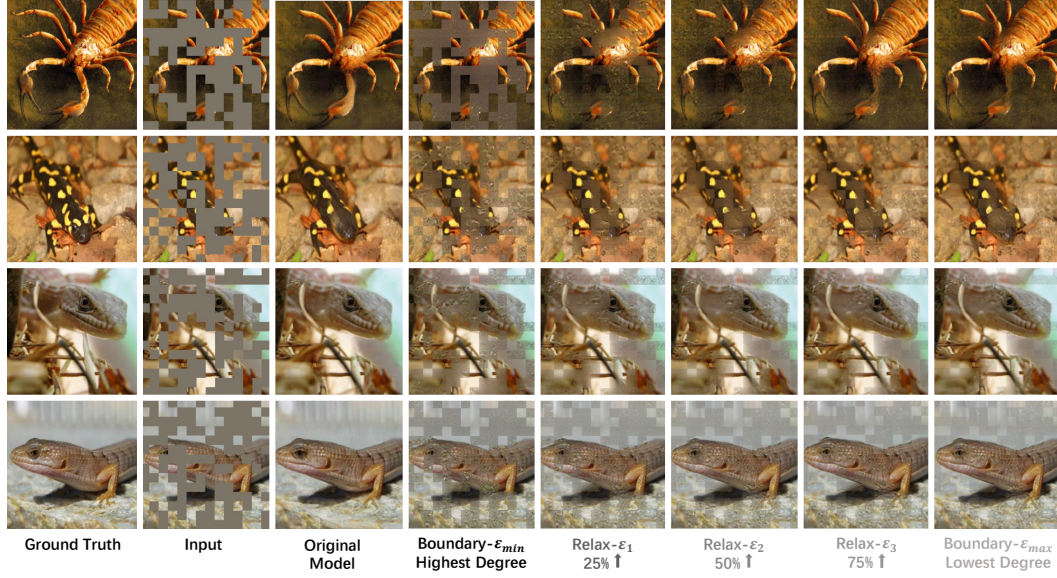
(a) Forget Set



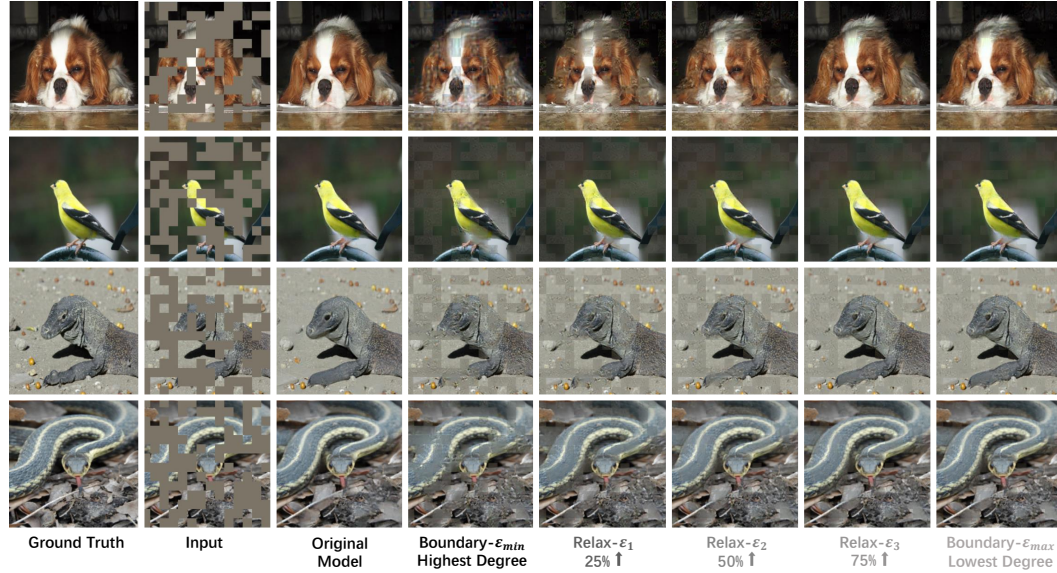
(b) Retain Set

Figure 8: VQ-GAN: generated images of cropping 50% at the center of the image under different degrees of unlearning completeness requirements. The upper half (a) represents the forget set, and the lower half (b) represents the retain set. Our method first determines the two boundary conditions of unlearning, and then linearly increases the value of ϵ within its range (here, we increase by 25% each time) to adjust the balance between unlearning completeness and model utility.

MAE. We verify the control effect of our controllable unlearning framework within the reconstruction task using the MAE. The results in Figure 9 indicate that our method can effectively control the completeness of unlearning in image reconstruction tasks as well.



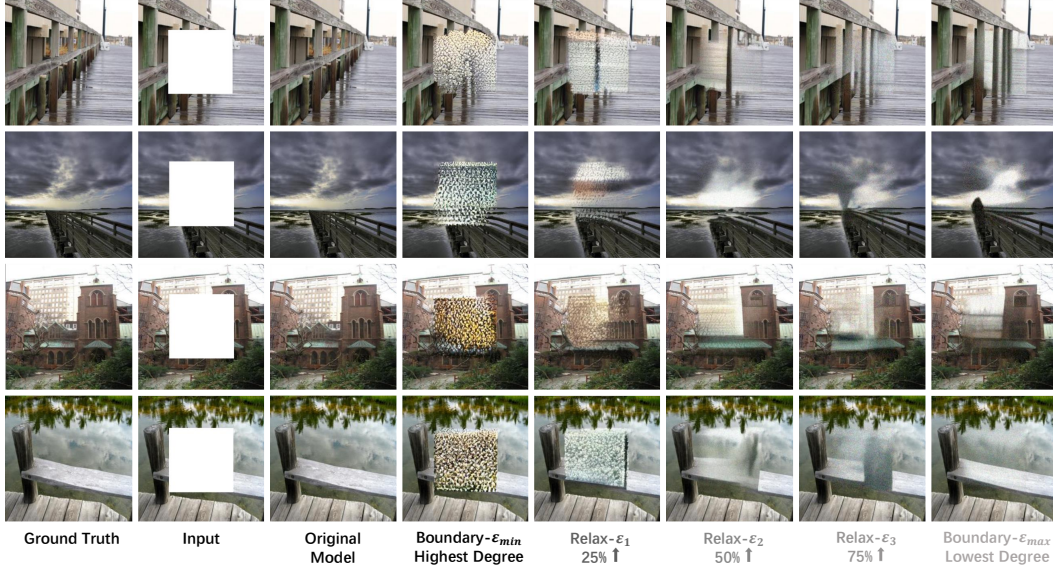
(a) Forget Set



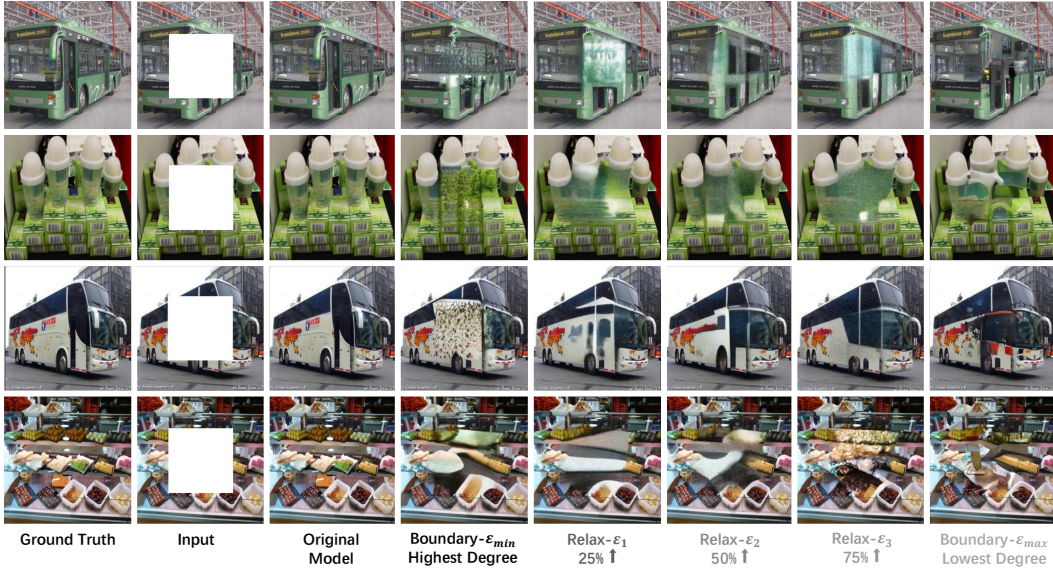
(b) Retain Set

Figure 9: MAE: construction of random masked images under different degrees of unlearning completeness requirements. We set the proportion of the random mask to 50%. The upper half (a) represents the forget set, and the lower half (b) represents the retain set. Our method first determines the two boundary conditions of unlearning, and then linearly increases the value of ϵ within its range (here, we increase by 25% each time) to adjust the balance between unlearning completeness and model utility.

Diffusion model. We validate the control effect of our controllable unlearning framework within the inpainting task of a diffusion model. As shown in Figure 10, the findings illustrate that our method can effectively adjust the balance between the completeness of unlearning and the utility of the model in the context of a diffusion model.



(a) Forget Set



(b) Retain Set

Figure 10: Diffusion model: generated images of cropping 50% at the center of the image under different degrees of unlearning completeness requirements. The upper half (a) represents the forget set, and the lower half (b) represents the retain set. Our method is also effective when applied to the diffusion model.

H ABLATION STUDY

To verify the robustness of our method on mainstream I2I generative models and various image generation tasks, we conducted the following ablation studies: i) we vary the cropping patterns to

demonstrate robustness across multiple image generation tasks; ii) we decrease the linear increment size of ε to validate that our method allows for more fine-grained control; and iii) we alter the cropping ratios to confirm the robustness of our method to changes in crop ratio.

H.1 MORE GENERATIVE TASKS

Similar to validating unlearning in classification models through Membership Inference Attacks Choi & Na (2023), generative models can also be assessed for unlearning robustness by employing attack methods to reconstruct the forget set. Although there is substantial research in this area Kumari et al. (2023); Petsiuk & Saenko (2025), it typically focuses on concept unlearning in text-to-image generative models. In contrast, our focus is on unlearning in image-to-image generative models. Unlike unlearning a single concept, our goal is to unlearn the influence of a set of samples or their distribution on the model. This makes it challenging to validate the effectiveness and robustness of our method through attacks. Specifically, we validate the effectiveness and robustness of our controllable unlearning framework for image extension tasks on VQ-GAN by varying the patterns of cropping. The results indicate that our controllable unlearning framework is robust to different cropping patterns.

H.1.1 OUTPAINTING TASK

We retain 25% of the image center and utilize VQ-GAN for image outpainting. As shown in Figure 11, our method produces outpainting on the forget set that is most similar to Gaussian noise, and the outpainting performance on the retain set shows the least decline compared to the original model.

H.1.2 UPWARD EXTENSION TASK

We crop the upper half of the image, retain the lower half, and employ VQ-GAN for image extension. The results in Figure 12 indicate that our method produces extension on the unlearning set that closely resembles Gaussian noise, and on the retain set, the extension performance decreases the least compared to the original model.

H.1.3 LEFTWARD EXTENSION TASK

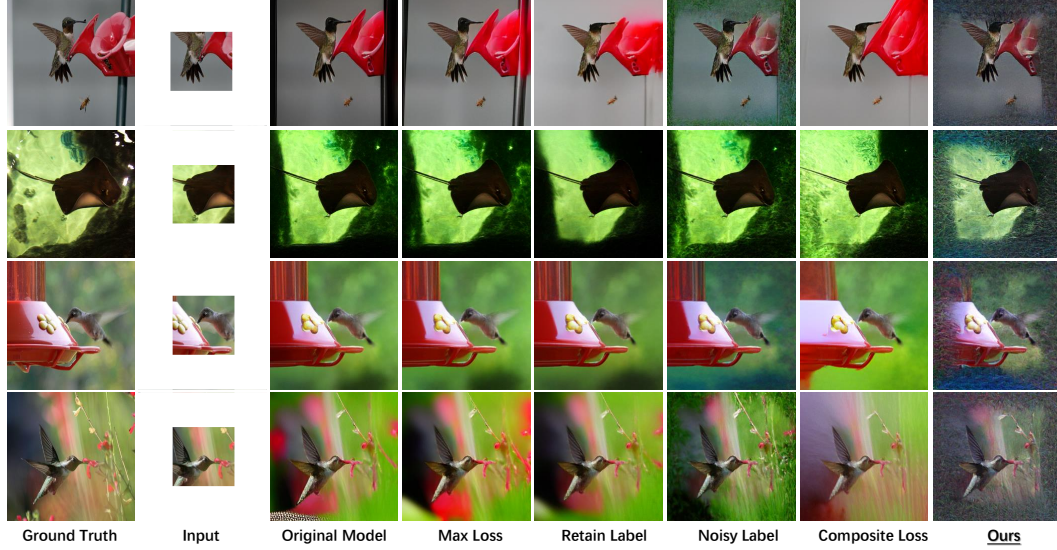
We crop the right half of the image, retain the left half, and use VQ-GAN for image extension. As shown in Figure 13, our method produces leftward extension on the forget set that closest resembles Gaussian noise and, on the retain set, the leftward extension performance exhibits the minimal decrease compared to the original model.

H.2 MORE FINE-GRAINED CONTROL OF UNLEARNING COMPLETENESS

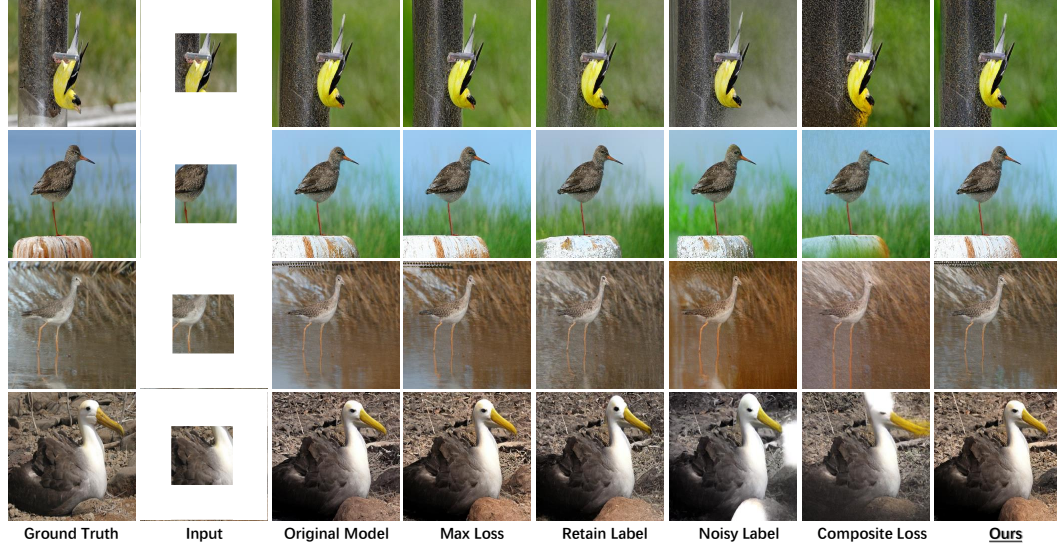
After obtaining two boundary points of unlearning, our controllable unlearning framework linearly increases within its valid range to balance the completeness of unlearning and the utility of the model. However, in the main paper, the increase of ε is by 25% each time. For example, if the range of ε is $[1, 9]$, then the sequence of ε values would be $\{3, 5, 7\}$. It is evident that the increments of ε are quite substantial, which results in a coarser granularity of control. Here, we reduce the linear increment of ε to extend the effectiveness of our controllable unlearning framework across various image generation tasks in VQ-GAN. The results show that our framework can achieve fine-grained control.

H.2.1 OUTPAINTING TASK

We retain the central 25% of the image and utilize VQ-GAN for image outpainting. The results in Figure 14 show that the performance of our controllable unlearning framework on the forget set gradually improves with the increase of ε , and the extent of decline in outpainting performance on the retain set, compared to the original model, is also reducing.

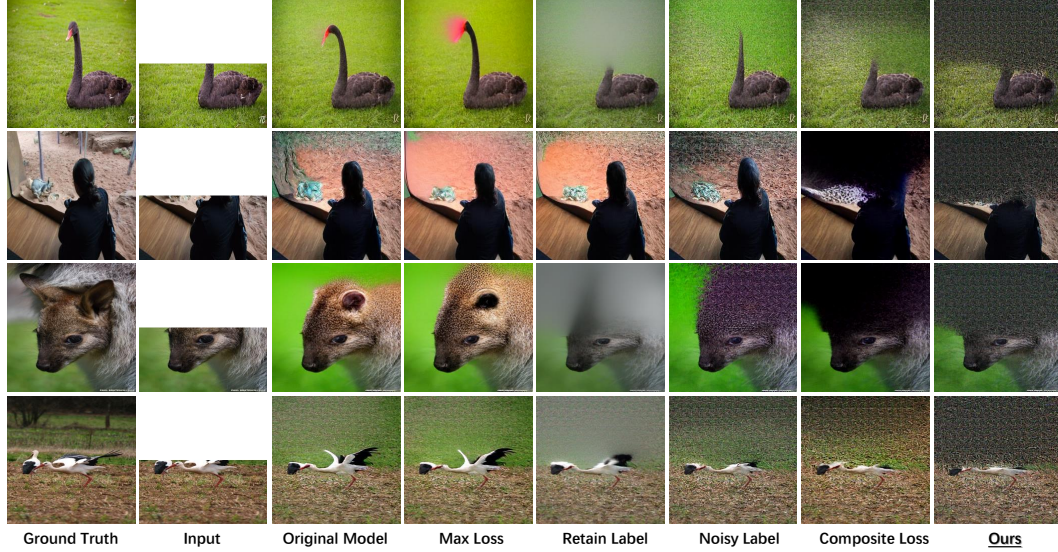


(a) Forget Set

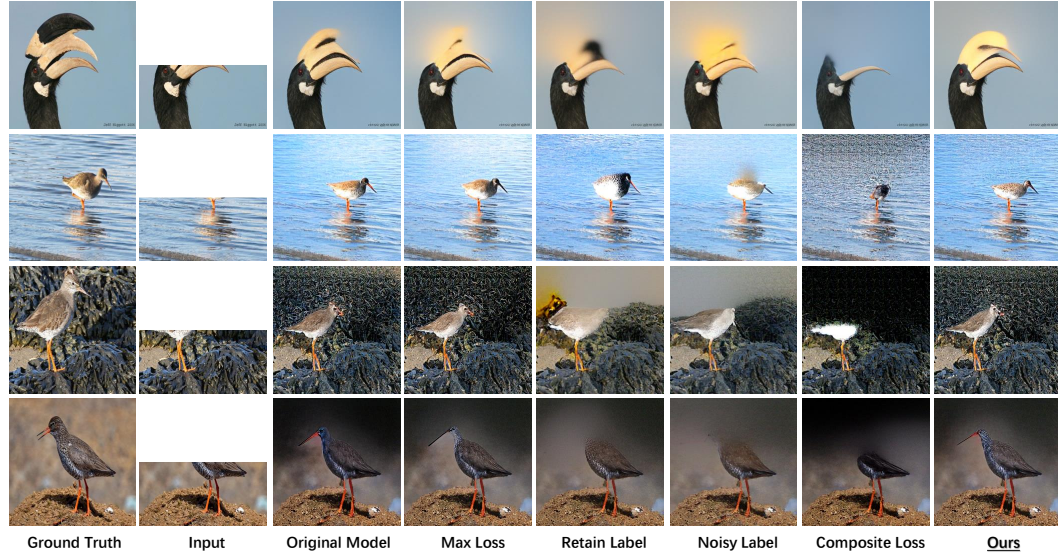


(b) Retain Set

Figure 11: Outpainting by VQ-GAN. We retain 25% of the image center. The upper half (a) designated as the unlearning set and the lower half (b) as the retain set. For each subset, we compared the performance of both the baselines and our method on the outpainting task, where "Ours" represents the boundary condition of unlearning in Phase I, indicating the point of highest degree of unlearning completeness. The results show that our method significantly outperforms the baselines on the outpainting task.

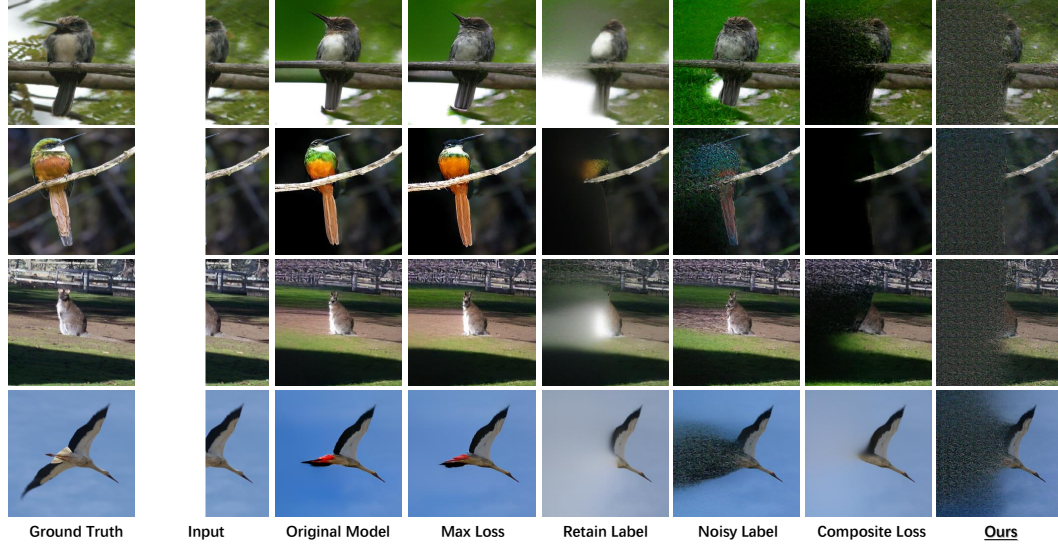


(a) Forget Set

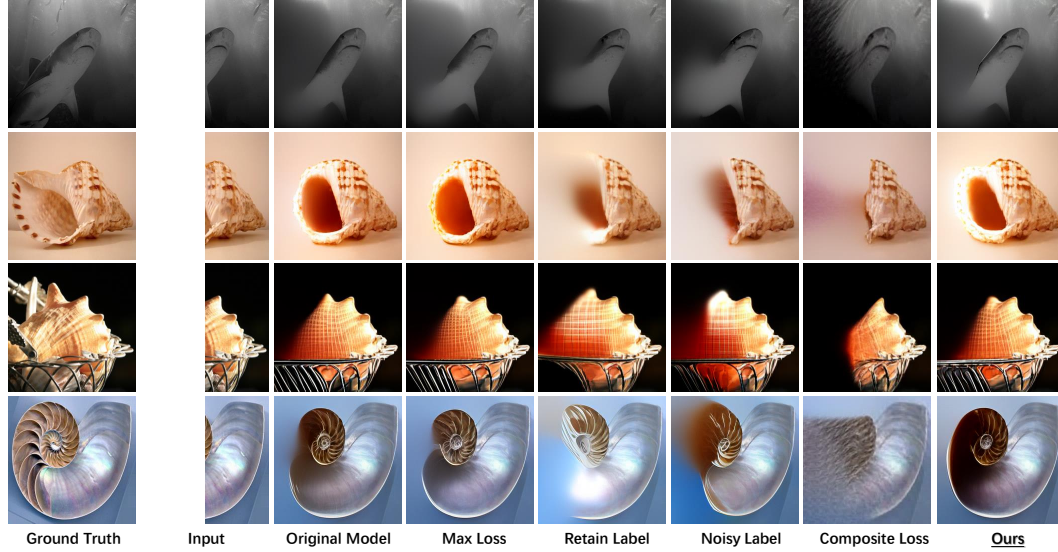


(b) Retain Set

Figure 12: Upward extension by VQ-GAN. We retain 50% of the lower half of the image. The upper half (a) is the forget set, and the lower half (b) is the retain set. For each set, we compare the performance of the baselines and our method on the upward extension task, where "Ours" represents the unlearning boundary condition in Phase I, which is the point of the highest degree of unlearning completeness. The results suggest that our method also significantly outperforms the baselines on the upward extension task.

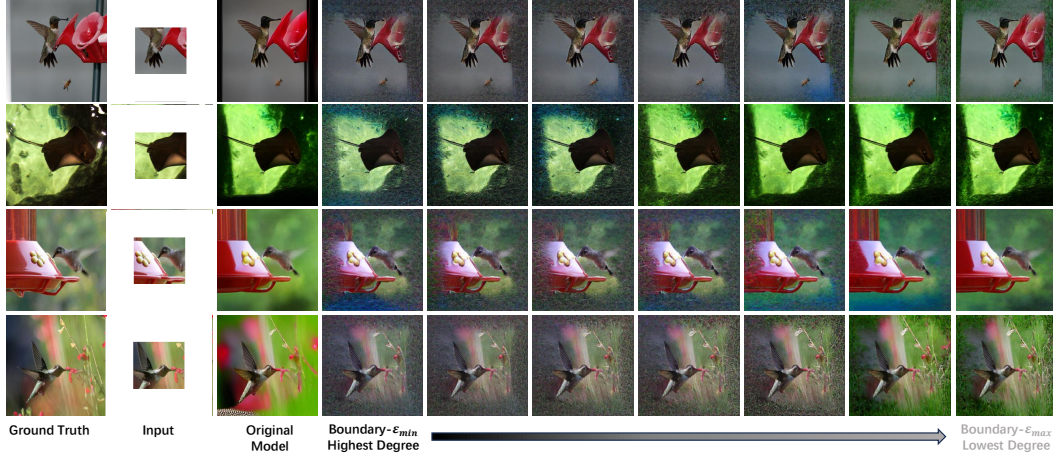


(a) Forget Set

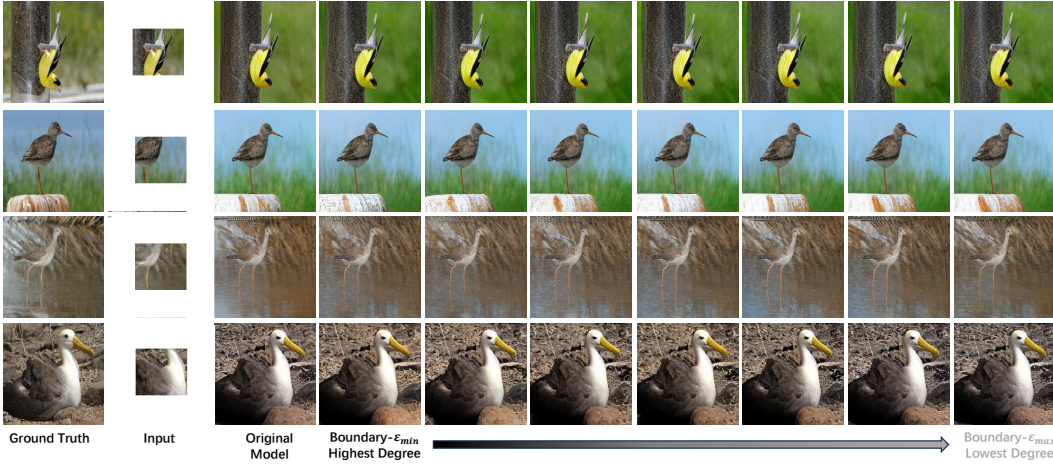


(b) Retain Set

Figure 13: Leftward extension by VQ-GAN. We retain 50% of the right half of the image. The upper half (a) is the forget set, and the lower half (b) is the retain set. For each set, we compare the performance of the baselines and our method on the upward extension task, where "Ours" represents the unlearning boundary condition in Phase I, which is the point of highest degree of unlearning completeness. The results suggest that our method also significantly outperforms the baselines on the upward extension task.



(a) Forget Set

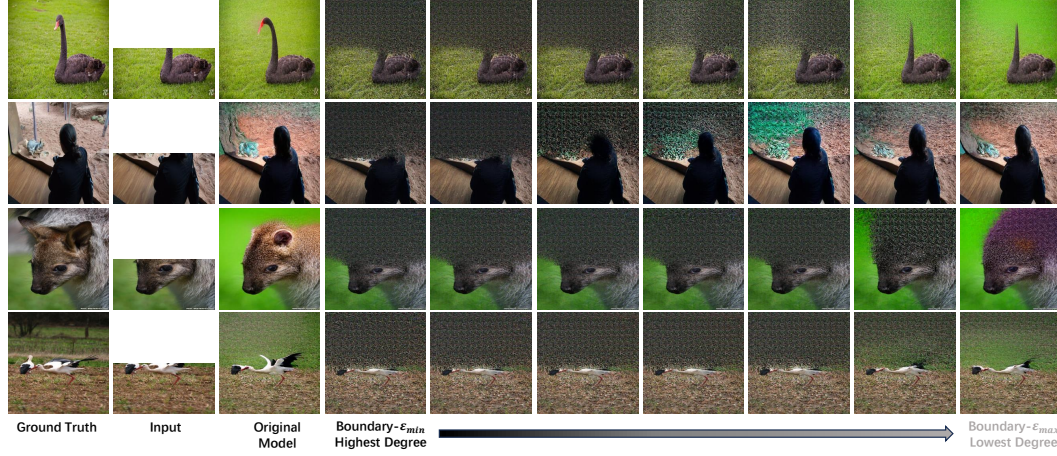


(b) Retain Set

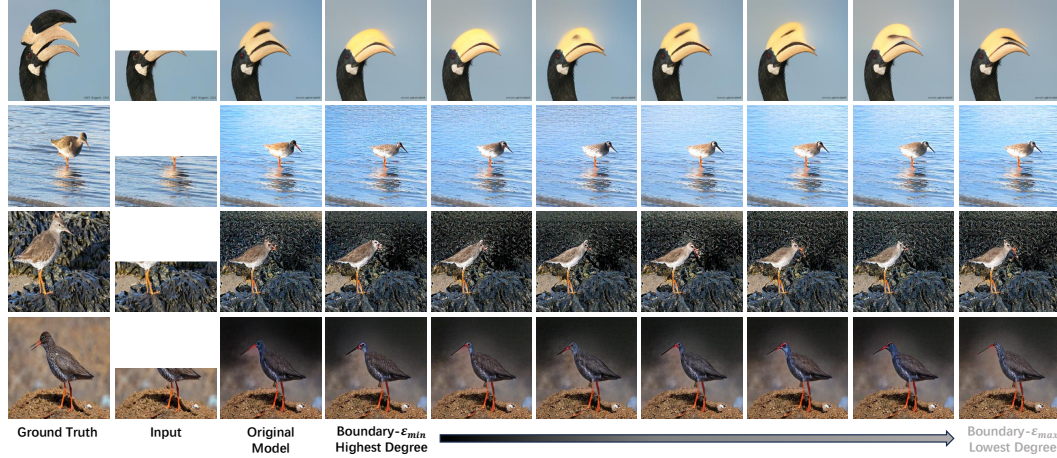
Figure 14: Outpainting by VQ-GAN under different degrees of unlearning completeness. We retain 25% of the image center. The upper half (a) is the forget set, while the lower half (b) is the retain set. For each part, we compare the unlearning effects of our method at different values of ϵ . "Highest" and "Lowest" represent the conditions of the highest and lowest degree of unlearning completeness, respectively. We increase ϵ 16% each time.

H.2.2 UPWARD EXTENSION TASK

We retain the lower half of the image center and crop the upper half, employing VQ-GAN for image extension. As shown in Figure 15, results indicate that, with an increase in the value of ϵ , the upward extension effectiveness on the forget set of our controllable unlearning framework gradually improves. Concurrently, the degree of decrease in upward extension effectiveness on the retain set, in comparison to the original model, also diminishes.



(a) Forget Set

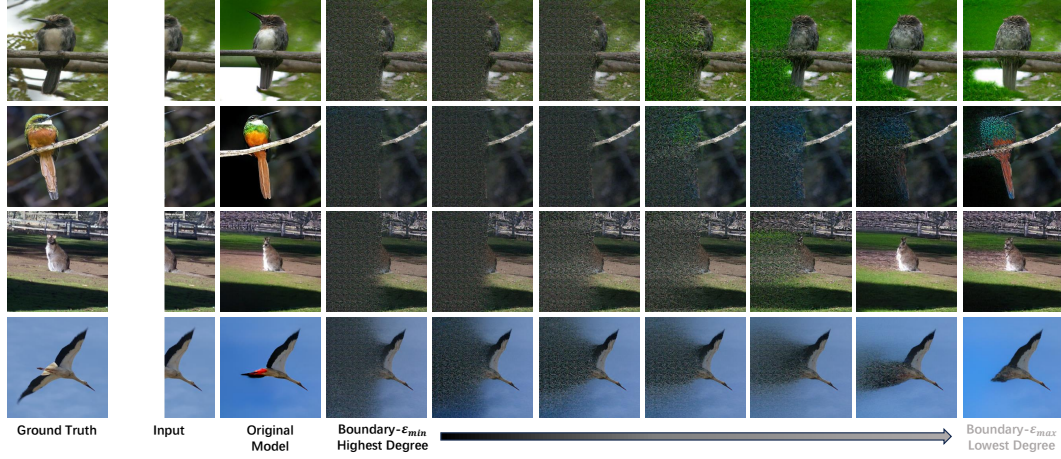


(b) Retain Set

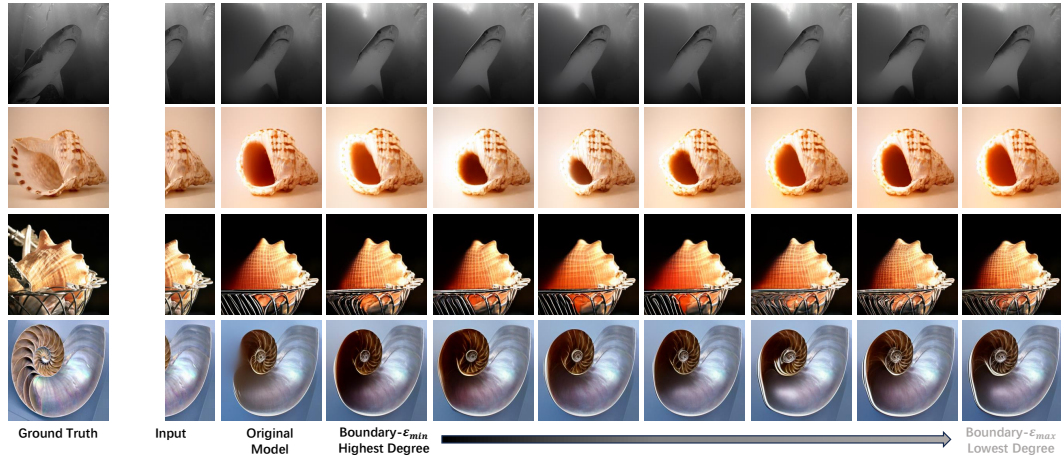
Figure 15: Upward extension by VQ-GAN under different degrees of unlearning completeness. We retain 50% of the lower half of the image. The upper half (a) is the forget set, while the lower half (b) is the retain set. For each part, we compare the unlearning effects of our method at different values of ϵ . "Highest" and "Lowest" represent the conditions of the highest and lowest degree of unlearning completeness, respectively. We increase ϵ 16% each time.

H.2.3 LEFTWARD EXTENSION TASK

We retain the right half of the image and utilize VQ-GAN to extend the image from the left. The results in Figure 16 demonstrate that the leftward extension performance on the forget set of our controllable unlearning framework progressively improves with the increase of ϵ , and the reduction in leftward extension performance on the retain set is also diminishing compared to the original model.



(a) Forget Set



(b) Retain Set

Figure 16: Leftward extension by VQ-GAN under different degrees of unlearning completeness. We retain 50% of the right half of the image. The upper half (a) is the forget set, while the lower half (b) is the retain set. For each part, we compare the unlearning effects of our method at different values of ϵ . "Highest" and "Lowest" represent the conditions of the highest and lowest degree of unlearning completeness, respectively. We increase ϵ 16% each time.

H.3 VARYING CROPPING PATTERNS AND RATIOS

In the preceding sections, we have demonstrated the performance of our controllable unlearning framework under various cropping patterns, yet the cropping ratio remained constant. By altering the cropping ratio on VQ-GAN, we validate the effectiveness of our controllable unlearning framework at different cropping ratios. The results indicate that our controllable unlearning framework is robust to different cropping ratios. Simultaneously, compared to larger cropping ratios, the extent of variation in the images generated under our controllable unlearning framework will be smaller for smaller cropping ratios.

H.3.1 INPAINTING TASK

We retain one-sixteenth of the image center and use VQ-GAN for image inpainting. The results in Figure 17 show that our controllable unlearning framework significantly outperforms the baselines in terms of unlearning effect on the forget set, most closely approximating Gaussian noise, and exhibits a lesser decline in unlearning effect on the retain set than the baselines. Simultaneously, we can finely control the balance between unlearning completeness and model utility.

H.3.2 DOWNWARD EXTENSION TASK

We crop the bottom 25% of the image and utilize VQ-GAN for image extension from the bottom. As shown in Figure 19, the results demonstrate that our controllable unlearning framework significantly surpasses the baselines in terms of the unlearning effect on the forget set, closely approximating Gaussian noise, and shows a lesser reduction in unlearning effect on the retain set compared to the baselines. At the same time, we can finely adjust the balance between unlearning completeness and model utility.

H.3.3 RIGHTWARD EXTENSION TASK

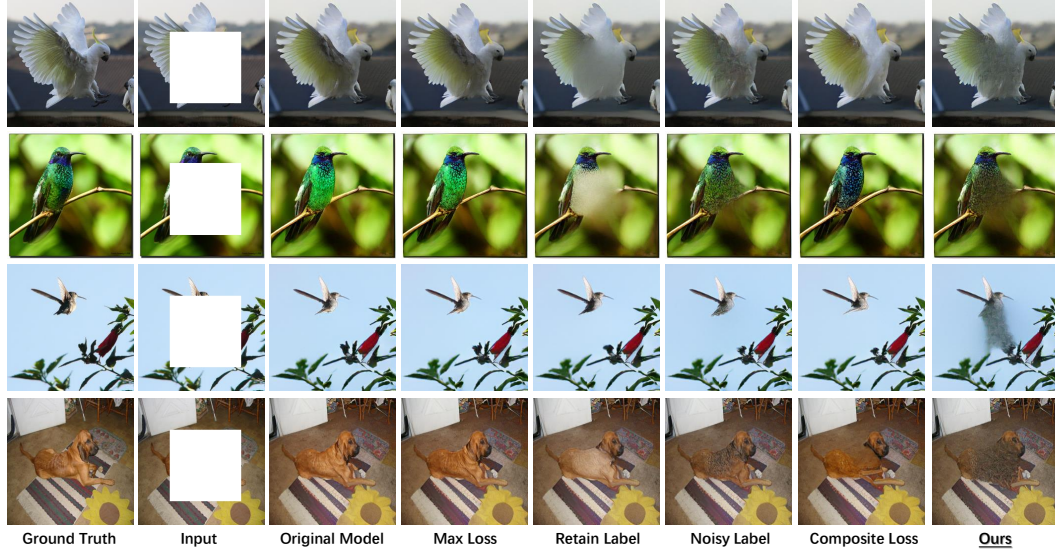
We crop the right 25% of the image and utilize VQ-GAN for image extension from the bottom. The results in Figure 21 demonstrate that our controllable unlearning framework significantly surpasses the baselines in terms of the unlearning effect on the forget set, closely approximating Gaussian noise, and shows a lesser reduction in unlearning effect on the retain set compared to the baselines. At the same time, we can finely adjust the balance between unlearning completeness and model utility.

I T-SNE ANALYSIS FOR CONTROLLABLE UNLEARNING

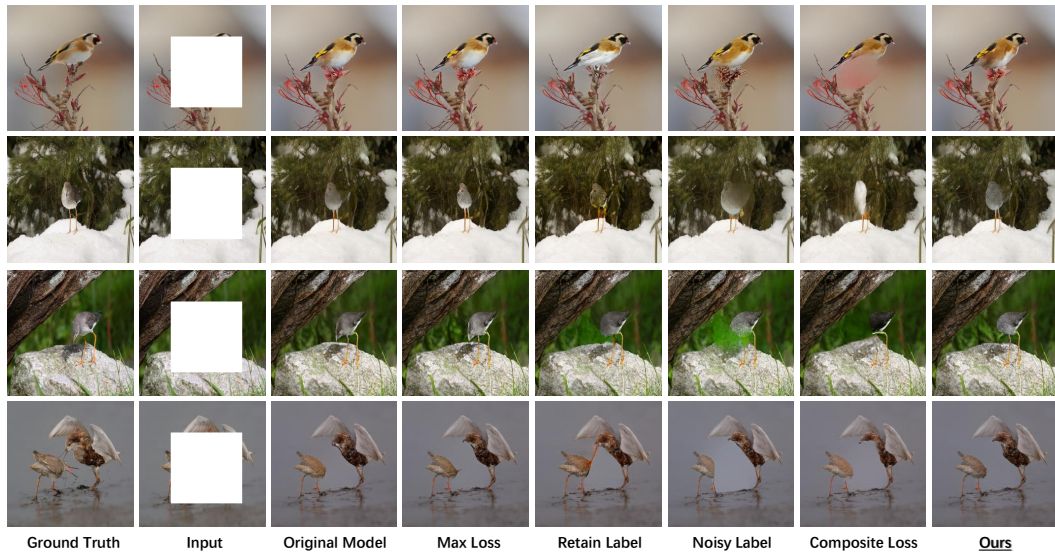
In Table 2 of the main paper, we present the evaluation metrics corresponding to different degrees of unlearning completeness solutions (i.e., IS, FID and CLIP) obtained by our controllable unlearning framework in mainstream I2I generative models. Here, we analyze the images generated at different degrees of unlearning completeness for each corresponding model. We use T-SNE analysis to compare the clip embedding distances between the images generated on the forget set and retain set and the ground truth images. As shown in Figure 23, for any model, under the highest degree of unlearning completeness, the distance between the clip embeddings of the images generated on the forget set by the unlearned model and the ground truth images is larger, while the distance on the retain set is smaller. Simultaneously, as ϵ increases, the distance between the clip embeddings of the images generated on the forget set by the unlearning model and the ground truth images gradually decreases (still significantly higher than the situation of the retain set), and the distance on the retain set also gradually decreases. Lastly, among these three mainstream I2I generation model structures, the effect of VQ-GAN is the most significant.

J EFFICIENCY EXPERIMENTS FOR CONTROLLABLE UNLEARNING FRAMEWORK

In the main paper, we analyze the convergence efficiency corresponding to different control functions $\psi(\theta)$ at each phase from a theoretical perspective, and based upon this analysis, we aim to enhance the unlearning efficiency of our controllable unlearning framework. Here, we validate our



(a) Forget Set

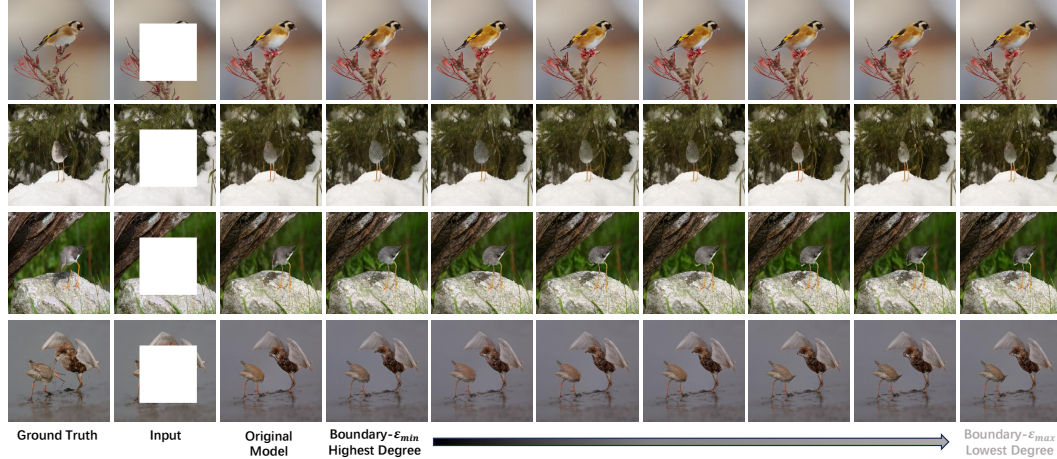


(b) Retain Set

Figure 17: Generated images of cropping 25% at the center of the image. We crop the center 1/16 of the image. The upper half (a) is the forget set, and the lower half (b) is the retain set. For each set, we compare the performance of the baselines and our method on the inpainting task, where "Ours" represents the extreme case of the unlearning boundary in Phase I, that is, the point of highest degree of unlearning completeness.

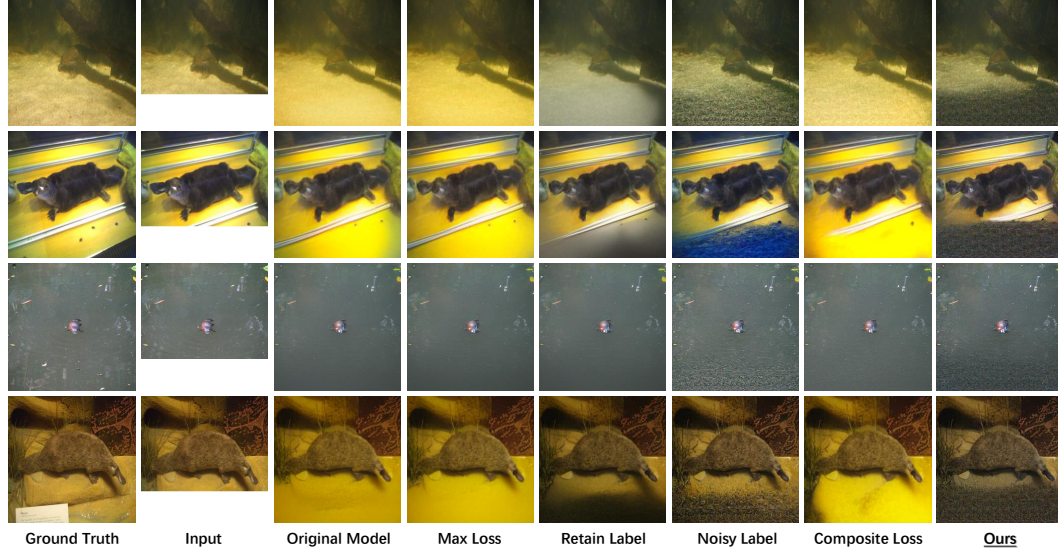


(a) Forget Set

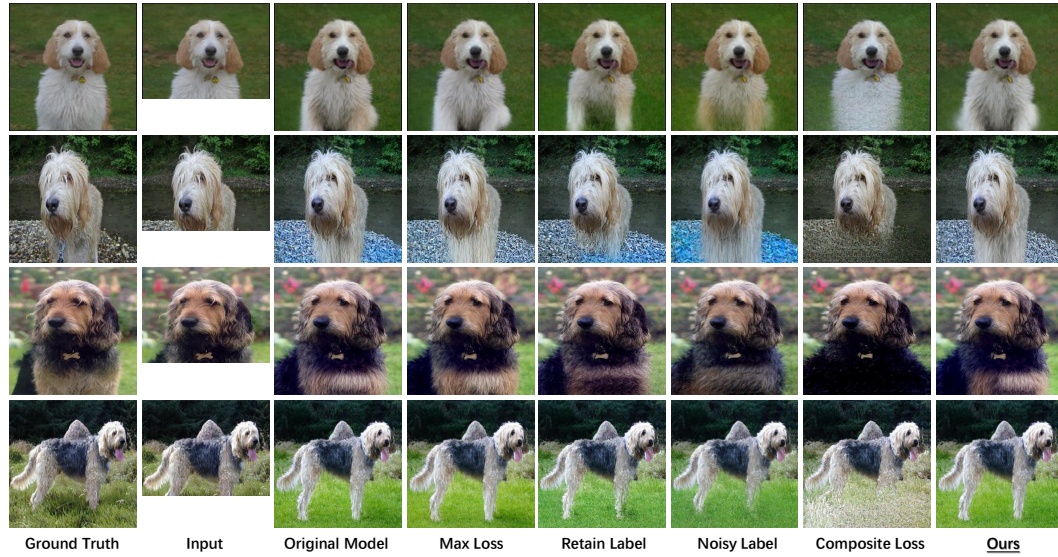


(b) Retain Set

Figure 18: Generated images of cropping 50% at the center of the image under different degrees of unlearning completeness requirements. We crop the central 1/16 of the image. The upper half (a) represents the forget set, and the lower half (b) represents the retain set. For each section, we compare the effectiveness of our method’s unlearning under different values of ϵ . Here, “Highest” and “Lowest” indicate the conditions of the highest and lowest degree of unlearning completeness, respectively.

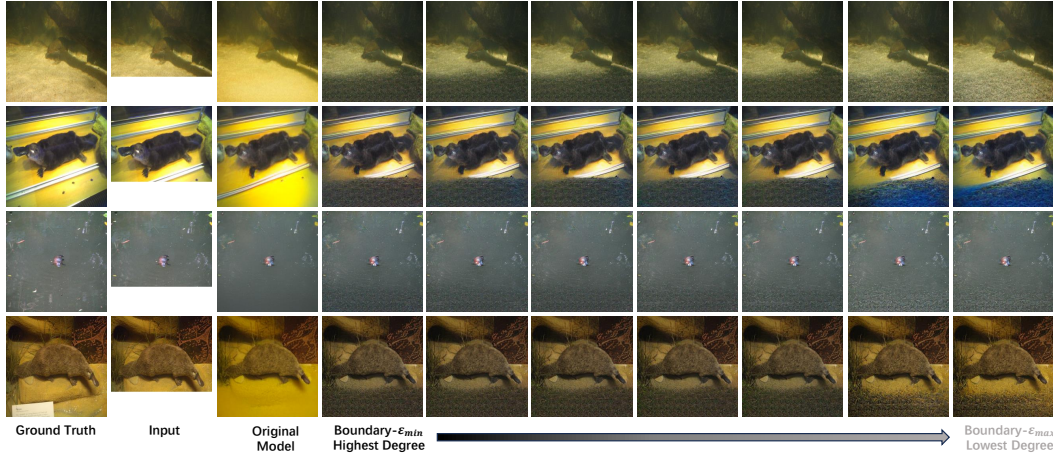


(a) Forget Set

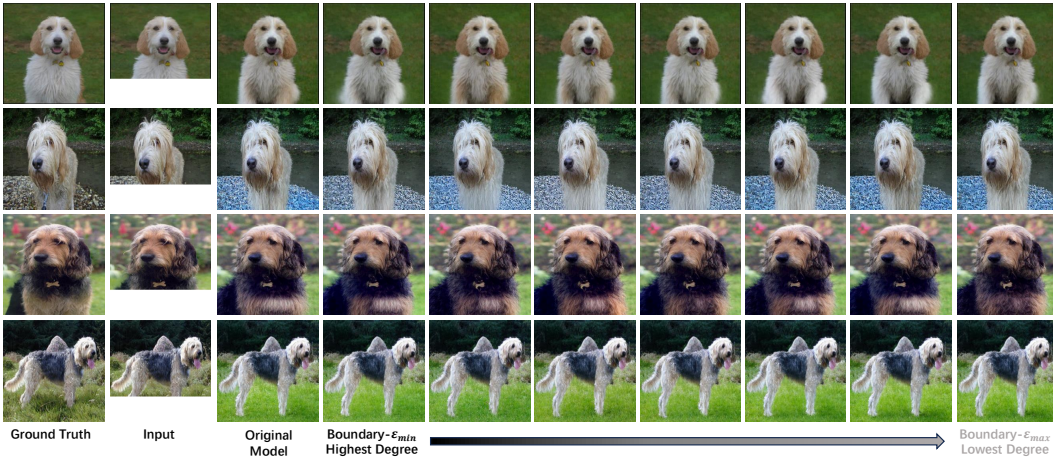


(b) Retain Set

Figure 19: Downward extension by VQ-GAN. We crop the bottom 25% of the image. The upper half (a) is designated as the forget set, and the lower half (b) as the retain set. For each section, we compared the performance of the baselines and our method on the downward extension task, where "Ours" denotes the unlearning boundary condition in Phase I, that is, the point of highest degree of unlearning completeness.

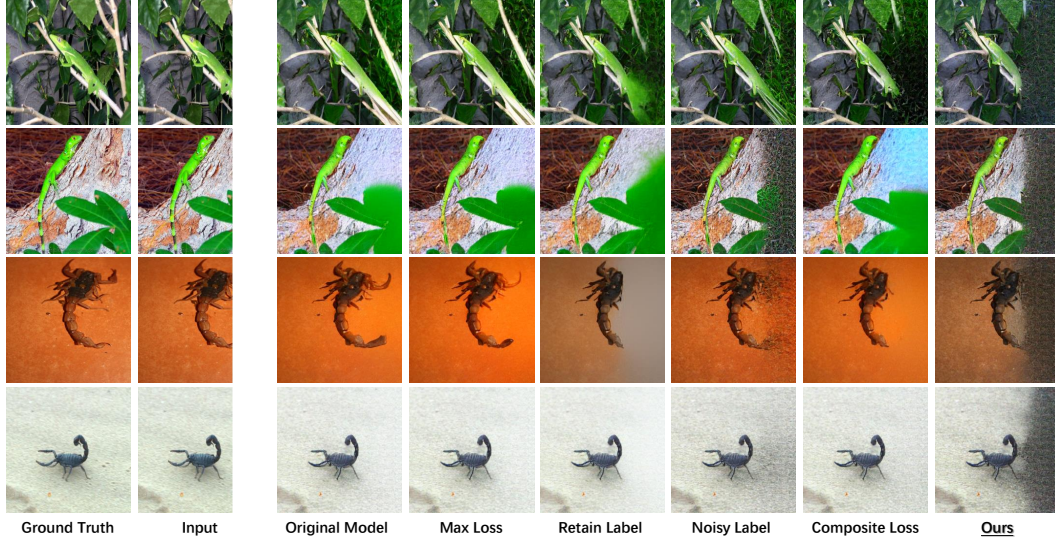


(a) Forget Set

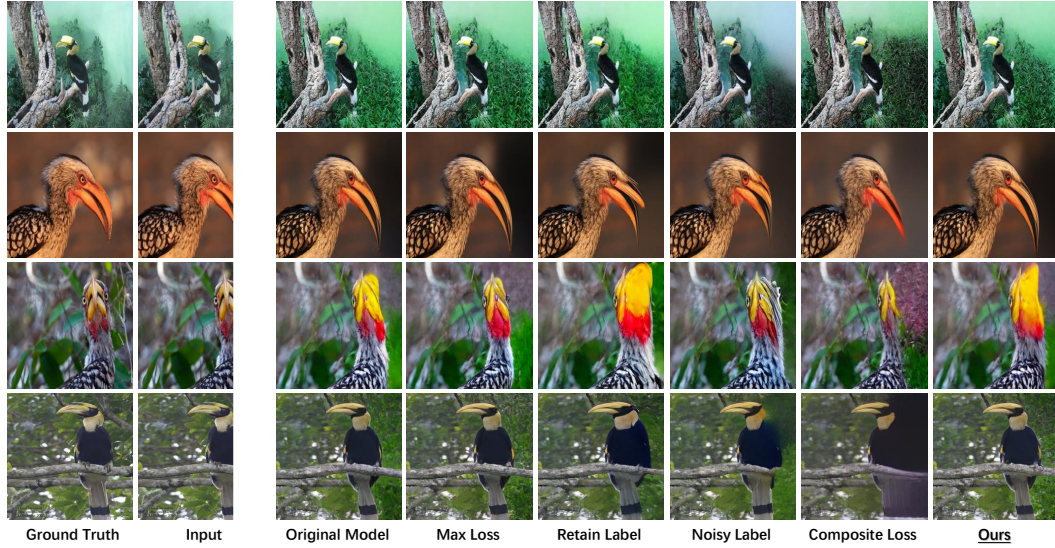


(b) Retain Set

Figure 20: Downward extension by VQ-GAN under different degrees of unlearning completeness. We crop the bottom 25% of the image. The upper half (a) represents the forget set, and the lower half (b) represents the retain set. For each section, we compare the effectiveness of our method’s unlearning under different values of ϵ . Here, “Highest” and “Lowest” indicate the conditions of the highest and lowest degree of unlearning completeness, respectively.

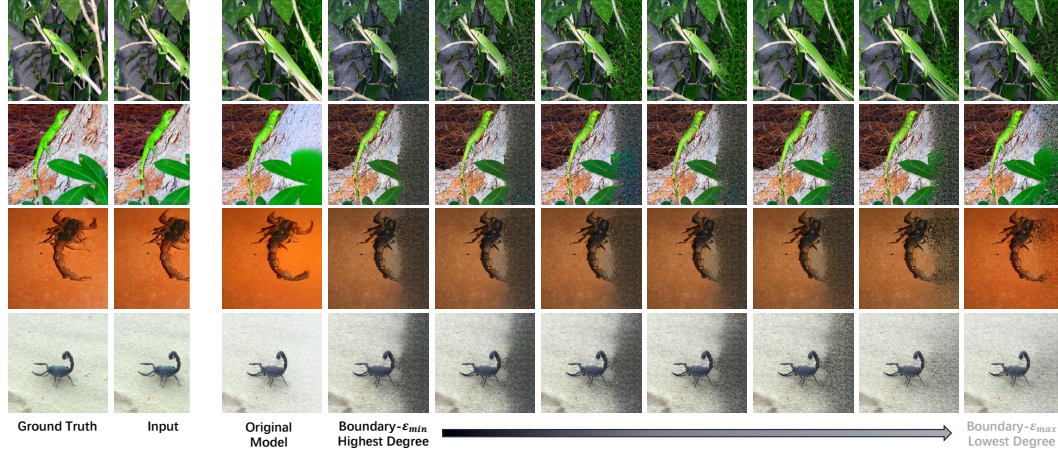


(a) Forget Set

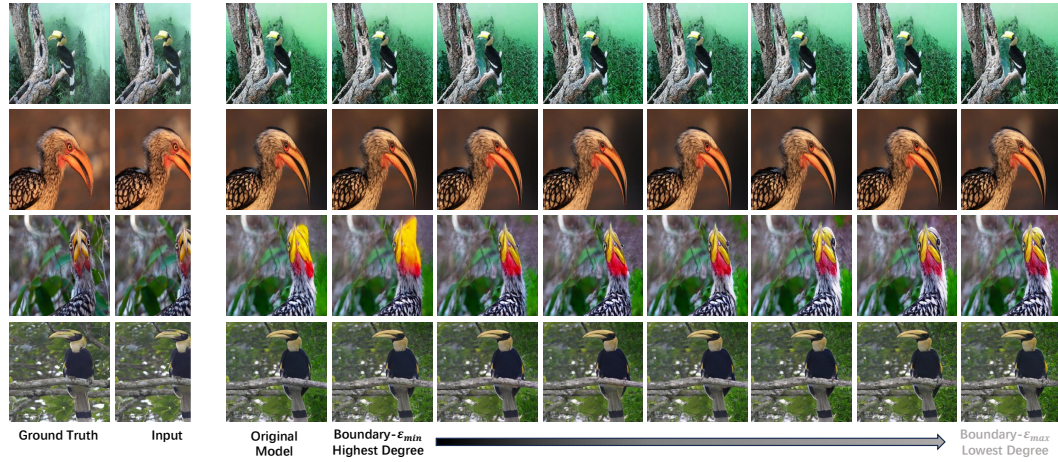


(b) Retain Set

Figure 21: Rightward extension by VQ-GAN. We crop the right 25% of the image. The upper half (a) is designated as the forget set, and the lower half (b) as the retain set. For each section, we compared the performance of the baselines and our method on the rightward extension task, where "Ours" denotes the unlearning boundary condition in Phase I, that is, the point of highest degree of unlearning completeness.



(a) Forget Set



(b) Retain Set

Figure 22: Rightward extension by VQ-GAN under different degrees of unlearning completeness. We crop the right 25% of the image. The upper half (a) represents the forget set, and the lower half (b) represents the retain set. For each section, we compare the effectiveness of our method’s unlearning under different values of ϵ . Here, “Highest” and “Lowest” indicate the conditions of the highest and lowest degree of unlearning completeness, respectively.

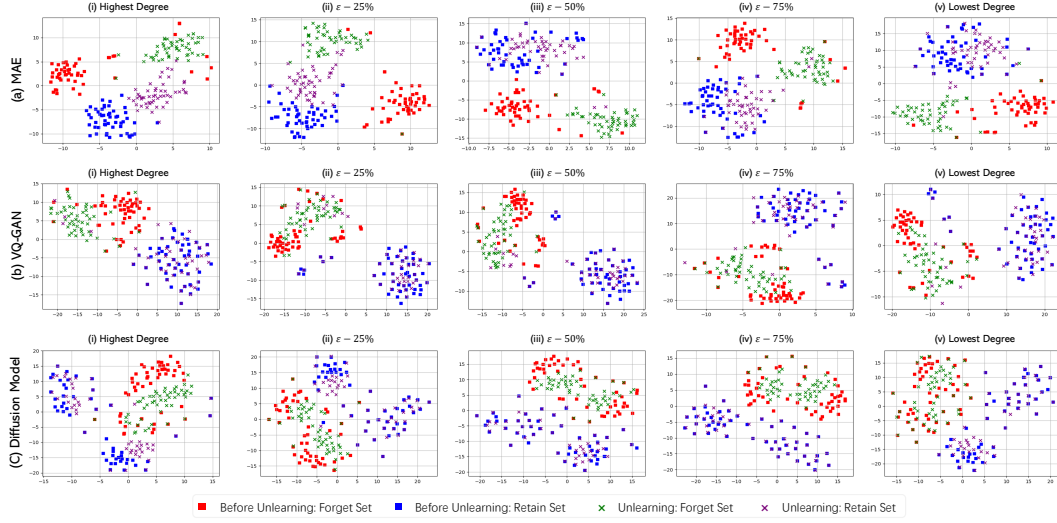


Figure 23: T-SNE analysis between images generated by our method and ground truth images under different degrees of unlearning completeness.

analysis on three mainstream I2I generative models. During the two different phases of controllable unlearning, we design the form of the control function $\psi(\theta)$ separately.

Specifically, in Phase I, we set $\psi(\theta) = \alpha \|\nabla f_1(\theta)\|^\delta$, where we test the convergence rates of $f_1(\theta)$ and $f_2(\theta)$, as well as the overall convergence rate, for $\delta = 1$, $\delta = 2$, $\delta = 3$, and $\delta = 4$. As shown in Figure 24, It is apparent that at Phase I for $c = 2$, that is $\psi(\theta) = \alpha \|\nabla f_1(\theta)\|^2$, the overall convergence rate is optimal.

In Phase II, we set $\psi(\theta) = \beta(f_1(\theta) - \varepsilon)^\delta$, where we tested the convergence rates for $\delta = 1$ and $\delta = 3$. Subsequently, we changed the form of $\psi(\theta)$ to $\psi(\theta) = \beta(f_1(\theta) - \varepsilon)^\delta \|\nabla f_1(\theta)\|^2$, and we tested the convergence rates for $\delta = 1$ and $\delta = 3$. Comparing the aforementioned scenarios, the overall optimal convergence rate in Phase II is obtained when $\psi(\theta) = \beta(f_1(\theta) - \varepsilon)^1 \|\nabla f_1(\theta)\|^2$.

You may include other additional sections here.

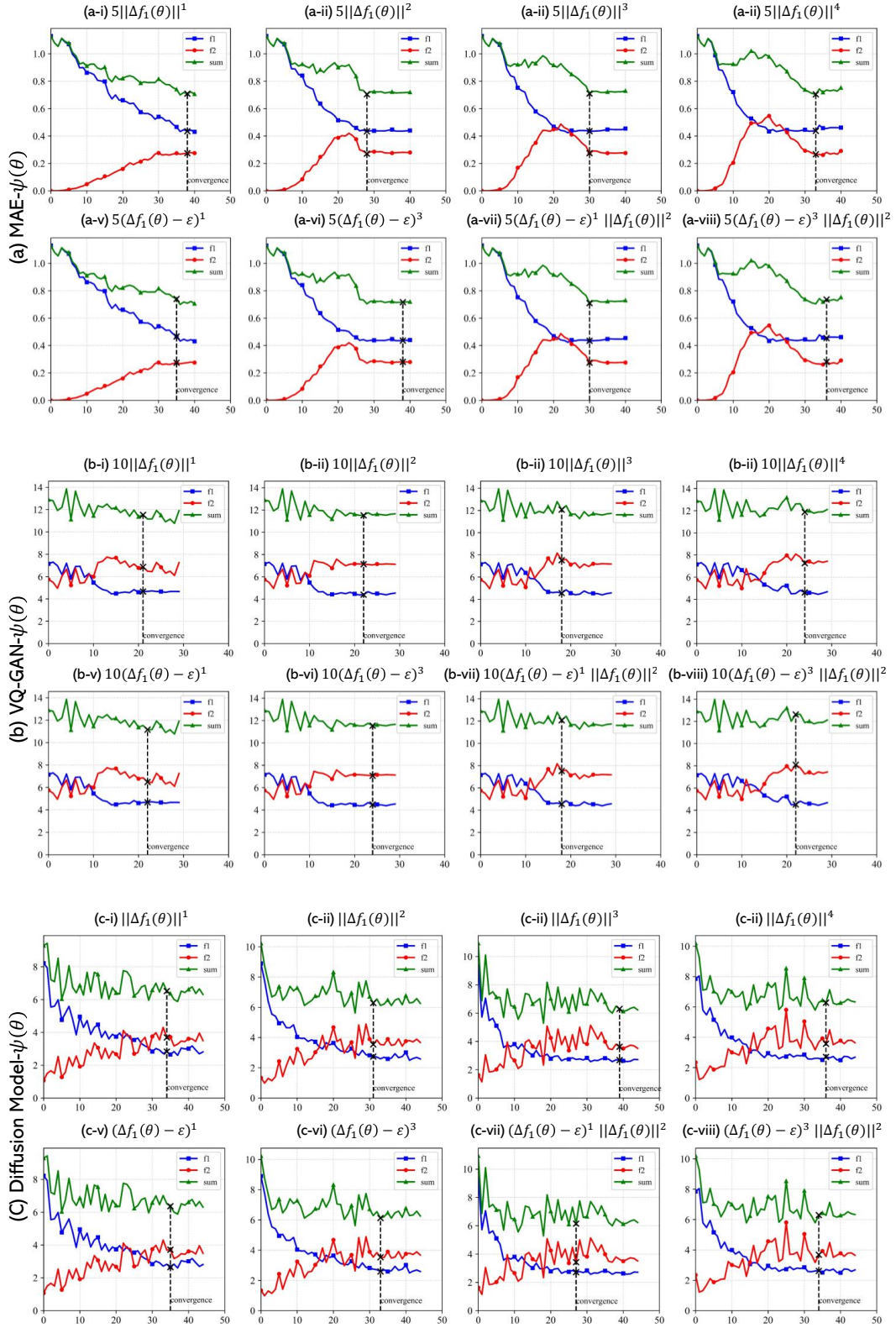


Figure 24: The convergence rates under different control functions $\psi(\theta)$. As illustrated in figure, include three sections: MAE, VQ-GAN, and the diffusion model. Each section contains two rows, corresponding to Phase I and Phase II, respectively. The titles on each subplot indicate the forms of the control function $\psi(\theta)$.