
Unlearning in Diffusion Models: A Unified Framework with KL Divergence and Likelihood Constraints

Shervin Khalafi¹ Alejandro Ribeiro¹ Dongsheng Ding²

Abstract

Unlearning in diffusion models aims to remove undesirable data or concepts while preserving the utility of pretrained models—two fundamentally conflicting objectives. We propose a principled constrained optimization framework that formulates unlearning as minimizing the deviation from a pretrained model, subject to explicit separation constraints from the unlearning distributions. Specifically, we formulate three constrained optimization problems based on reverse and forward KL divergences, and likelihood constraints. The first two generalize existing approaches for concept and data unlearning, while the third offers a novel and natural formulation for unlearning. Despite the nonconvexity of the KL constraints, we establish strong duality for all three problems, enabling us to explicitly characterize their optimal solutions as unlearning targets and develop primal–dual algorithms for each formulation. Experimental results demonstrate that our KL-constrained approach achieves superior retention-unlearning tradeoffs compared to weight-based baselines for concept and data unlearning, and that our likelihood-based approach matches unlearning effectiveness while better preserving retained concepts compared to baselines.

1. Introduction

Generative diffusion models have emerged as an effective approach for synthesizing high-quality images that resemble training data and capture underlying concepts. However, this expressiveness also raises significant safety and ethical concerns, as such models may generate inappropriate content, including copyrighted material (Carlini et al., 2023) or

¹University of Pennsylvania ²University of Tennessee, Knoxville. Correspondence to: Shervin Khalafi <shervink@seas.upenn.edu>, Dongsheng Ding <dongshed@utk.edu>.

harmful content (Schuhmann et al., 2022). While policymakers have made progress in regulating the removal of undesirable data or concepts from trained models (European Union, 2016; Goldman, 2020), it remains an open problem to develop effective technical methods to achieve this goal.

Recently, the classical concept of machine unlearning has been introduced in the context of diffusion models (Cao & Yang, 2015). A standard approach aims to retain the utility of pretrained models, ensuring high-quality image generation, while simultaneously preventing the generation of specific harmful data or concepts—two inherently *conflicting* objectives. This principle has motivated recent empirical advances, including concept unlearning (Gandikota et al., 2023; Feng et al., 2024) and data unlearning (Park & Park, 2025; Alberti et al., 2025). Often, these works employ the simple practice of balancing two conflicting objectives by combining them with weighted sums. However, such weight-based methods are inherently ad hoc, as the weights are treated as hyperparameters, making the approach sensitive and poorly generalizable across scenarios.

In this work, we take a principled constrained optimization approach to address the conflicting objectives in unlearning. To do so, we view retaining the utility of trained models as minimizing the distance between the trained model and the pretrained model, while preventing harmful generation is formulated as additional constraints that push the trained model away from specific data or concept distributions. As a result, the tradeoff between these objectives can be systematically characterized by a Lagrangian formulation. To measure the distance between two models, we employ reverse and forward KL divergences, which broadly correspond to concept and data unlearning, respectively. We summarize our key contributions in three aspects below.

- (i) **Problem Formulation.** We formulate a distribution optimization problem that minimizes the reverse KL divergence to a pretrained model, while constraining the reverse KL divergences to the distributions associated with the concepts to be unlearned to remain above user-specified thresholds. Formulation (RU) corresponds to unlearning at the concept level. We then consider an analogous optimization problem based on the forward KL divergence (FU), which corresponds

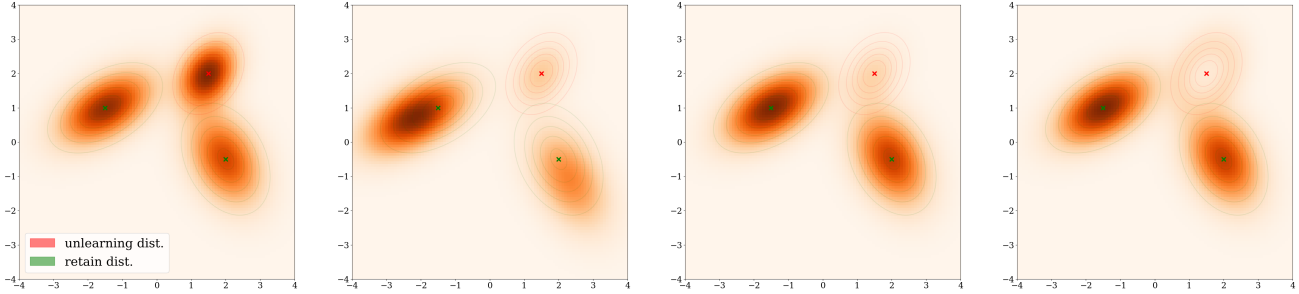


Figure 1. Heat maps of a three-Gaussian mixture before/after unlearning one mode (Right up). From left to right: pretrained model, reverse KL-constrained unlearning, forward KL-constrained unlearning, and likelihood-constrained unlearning.

to unlearning at the data level. Finally, we formulate a distribution optimization problem (LU) that constrains the expected likelihood of the concepts to be unlearned to remain below user-specified thresholds. This constrained framework extends the unlearning objectives, thereby revealing their commonalities and differences; see Section 1.1 for exact formulations and discussion.

(ii) **Theoretical Analysis.** Although the reverse and forward KL unlearning problems (RU) and (FU) are *non-convex*, we exploit the convexity of the image of a nonatomic vector measure to prove that they are strongly dual, rather than relying on standard convex duality. In contrast, strong duality for the likelihood unlearning problem (LU) directly follows from its convexity. In all cases, we explicitly characterize the optimal solutions, which serve as principled unlearning targets; see Figure 1 for visualizations and discussion. A key implication of strong duality is that the unlearning problems can be solved equivalently in the dual domain. This enables our development of primal–dual algorithms instantiated with diffusion models for each formulation.

(iii) **Empirical Results.** We demonstrate the effectiveness of our constrained optimization approach for concept and data unlearning in a text-to-image diffusion model. In concept unlearning, our method achieves the same degree of unlearning while deviating less from the pretrained model compared to equal-weights approaches. In data unlearning, assigning the same weight to every sample to be unlearned leads to excessive deviation from the pretrained model, whereas our method learns optimal weights that ensure unlearning while minimizing deviation. Furthermore, we show that our novel likelihood-based unlearning method achieves a better tradeoff between unlearning the target concept and preserving the retained concept. This allows us to perform strong unlearning without deviating significantly from the parts of the distribution we wish to retain.

1.1. Summary of our problem formulations

Within the optimization framework over the probability measure space Δ , we outline three principled unlearning targets and defer their exact characterizations to Section 2. Here, $q \in \Delta$ denotes the data distribution or the distribution induced by the pretrained model to be retained, and $q_u^i \in \Delta$ denotes the i th concept or data distribution to be unlearned.

Reverse KL-Constrained Unlearning. We minimize the reverse KL divergence between the model p and the reference q to ensure that the model is *close* to the reference. To remove the m undesirable concepts, we impose m reverse KL divergence constraints, each encouraging the model to *stay away from* an undesirable concept encoded by q_u^i . To ensure this separation, we enforce a lower bound b_i on each KL divergence by formulating a reverse KL-constrained distribution optimization problem,

$$\begin{aligned} & \underset{p \in \Delta}{\text{minimize}} && D_{\text{KL}}(p \parallel q) \\ & \text{subject to} && D_{\text{KL}}(p \parallel q_u^i) \geq b_i \text{ for } i \in [m]. \end{aligned} \quad (\text{RU})$$

Problem (RU) aims to minimize deviation from the pretrained model q —thereby maximizing model utility—while forgetting undesirable concepts. In Section 2.1, we capture a solution to Problem (RU) as a ratio of distributions,

$$p_{\text{rev}}^*(\cdot) \propto \frac{(q(\cdot))^{\alpha_0}}{\prod_{i=1}^m (q_u^i(\cdot))^{\alpha_i}}$$

where $\{\alpha_i\}_{i=0}^m$ is a set of non-negative exponents that balance the relative importance of each model. The subscript *rev* stands for reverse KL. The solution p_{rev}^* reflects unlearning by reducing the probability of sampling in regions where $q_u^i(\cdot)$ is large. This idea generalizes the concept unlearning target in concept erasing (Gandikota et al., 2023) and negation generation (Du et al., 2023).

Forward KL-Constrained Unlearning. We employ the forward KL (rather than reverse KL) divergence to measure the closeness between two distributions. To do so, we obtain

a variant of Problem (RU) in terms of forward KL,

$$\begin{aligned} & \underset{p \in \Delta}{\text{minimize}} && D_{\text{KL}}(q \| p) \\ & \text{subject to} && D_{\text{KL}}(q_u^i \| p) \geq b_i \text{ for } i \in [m]. \end{aligned} \quad (\text{FU})$$

Since the forward KL divergence is an expectation over the data distribution, it essentially optimizes the log-likelihood of the data under the model—often approximated by the evidence lower bound in diffusion models (Lai et al., 2025). Therefore, Problem (FU) seeks to maximize the likelihood of the data to be retained while reducing the likelihood of the data to be forgotten. In Section 2.2, we characterize a solution to Problem (FU) as a difference of distributions,

$$p_{\text{fw}}^*(\cdot) \propto q(x) - \sum_{i=1}^m \alpha_i q_u^i(\cdot)$$

where $\{\alpha_i\}_{i=1}^m$ is a set of non-negative weights that balance the relative importance of each distribution. The subscript fw stands for forward KL. The log-likelihood reduction of forgetting data is often referred to as data unlearning in the literature (Feng et al., 2024; Alberti et al., 2025). Formulation (FU) is distinguished by its weaker assumption: it requires only sample access to the unlearning distributions q_u^i , whereas the reverse KL and likelihood formulations assume access to the associated diffusion models.

Likelihood-Constrained Unlearning. In addition to the nonconvex reverse KL divergence constraints in Problem (RU), we propose an alternative formulation that constrains the likelihood of the unlearning data under the model,

$$\begin{aligned} & \underset{p \in \Delta}{\text{minimize}} && D_{\text{KL}}(p \| q) \\ & \text{subject to} && \mathbb{E}_p[q_u^i] \leq \epsilon_i \text{ for } i \in [m]. \end{aligned} \quad (\text{LU})$$

Problem (LU) aims to minimize the deviation from the pretrained model q while directly reducing the likelihood of the unlearning concepts appearing in the generated samples. In Section 2.3 we capture a solution to Problem (LU),

$$p_{\text{revl}}^*(\cdot) \propto \frac{q(\cdot)}{e^{\sum_{i=1}^m \alpha_i q_u^i(\cdot)}}$$

where $\{\alpha_i\}_{i=1}^m$ is a set of non-negative weights that balance the relative importance of each model. The subscript revl stands for reverse KL with likelihood. The exponential term downweights samples that have high likelihood under the unlearning distributions. Compared with the framework of constrained relative-entropy minimization (Koyejo & Ghosh, 2013), our formulation chooses the moment functions to be unlearning-distribution likelihoods, thereby turning the exponential tilt into a principled mechanism for pushing the learned distribution away from undesired concepts.

To illustrate, we use a three-Gaussian mixture example to compare the three unlearning solutions, as shown in Figure 1. While the formulations (RU) and (LU) both reduce

the probability of sampling from the unlearning concepts, their resulting solutions differ significantly. In particular, reverse KL-constrained unlearning achieves this by pushing the other modes we wish to retain away from the unlearned concept, whereas likelihood-constrained unlearning does not. This occurs because, for a sample x where q_u^i is small, the solution p_{revl}^* is more strongly altered by q_u^i than p_{revl}^* (proportional vs exponential).

Background on Diffusion Models. A diffusion model consists of two stochastic processes: a forward process and a backward process. The forward process starts from a data point x_0 and gradually adds noise according to $x_t = \sqrt{1 - \beta_t}x_{t-1} + \sqrt{\beta_t}w_t$ for $t \in [T]$, where $\beta_t \in (0, 1)$ is the stepsize and $w_t \sim \mathcal{N}(0, I)$ is the Gaussian noise. The backward process starts from pure noise $x_T \sim \mathcal{N}(0, I)$ and iteratively removes noise. For instance, the DDPM sampler updates $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t + \eta_t s(x_t, t) + \sigma_t z_t)$, where $\alpha_t = 1 - \beta_t$, $\eta_t > 0$ is the stepsize, σ_t is the noise scaling, and $z_t \sim \mathcal{N}(0, I)$ is the Gaussian noise. A crucial component of the backward process is the score function $s(x_t, t): X \times [T] \rightarrow \mathbb{R}^d$; when it is specified accurately, the backward process converges to the data distribution (Chen et al., 2023; Li & Yan, 2025). Throughout the paper, we focus on score-based diffusion models, denoting each probability measure p by its score function $s_p: X \times [T] \rightarrow \mathbb{R}^d$. We use $p_0(\cdot; s_p)$ to denote the marginal distribution of the terminal sample x_0 , and $p_{0:T}(\cdot; s_p)$ to denote the path distribution of the trajectory $x_{0:T}$, both induced by the score function s_p . We use analogous notation for the pretrained models q and q^i .

2. Unlearning in Distribution Space

In Sections 2.1–2.3, we formally present, in distribution space, the three formulations presented in Section 1.1. We prove strong duality in all three cases, despite the nonconvexity of the KL-divergence constraints.

2.1. Reverse KL-Constrained Unlearning

We begin with Problem (RU). Let $q: \mathcal{B}(X) \rightarrow [0, 1]$ be the probability measure induced by a pretrained model, where $\mathcal{B}(X)$ is a Borel σ -algebra over a bounded domain $X \subseteq \mathbb{R}^d$. In addition to the pretrained model, we consider a set of m probability measures $\{q_u^i: \mathcal{B}(X) \rightarrow [0, 1]\}_{i=1}^m$, each modeling a concept that we wish the model to forget/unlearn.

To ensure well-posedness, we assume that the probability measures of interest satisfy certain admissibility conditions.

Assumption 1 (Admissibility). *There exists a probability measure p satisfying two conditions: (i) p is absolutely continuous with respect to q, q_u^i , and (ii) p is bounded, i.e.,*

$$p \ll q, q_u^i \text{ for } i \in [m] \text{ and } p(x) < \infty \text{ for } x \in X. \quad (1)$$

The absolute continuity ensures the boundedness of reverse

KL divergence. It is practical to only consider bounded models, since diffusion models are represented by conditional Gaussian distributions. Problem (RU) is a *nonconvex* optimization problem because the superlevel set of reverse KL divergence is not convex. As a result, we cannot invoke the convexity of the feasible set to conclude strong duality from convex duality analysis (Boyd & Vandenberghe, 2004).

Assumption 2 (Feasibility). *There exists an admissible probability measure $p \in \Delta$: $D_{\text{KL}}(p \| q_u^i) > b_i$ for $i \in [m]$.*

Since q_u^i is focal on a specific concept or class, we can always construct p such that most of its mass is located on samples with low probability under q_u^i . This makes Assumption 2 practically achievable. Let p_{rev}^* be a solution to Problem (RU) and denote $P_{\text{rev}}^* = D_{\text{KL}}(p_{\text{rev}}^* \| q)$. Denote $\lambda := [\lambda_1, \dots, \lambda_m]^\top$. Let the Lagrangian for Problem (RU) be

$$L_{\text{rev}}(p, \lambda) = D_{\text{KL}}(p \| q) + \sum_{i=1}^m \lambda_i (b_i - D_{\text{KL}}(p \| q_u^i))$$

and the associated dual function is given by $D_{\text{rev}}(\lambda) = \min_{p \in \Delta} L_{\text{rev}}(p, \lambda)$, where the minimization is achieved at $p_{\text{rev}}^*(\cdot; \lambda)$. Let an optimal dual variable be $\lambda_{\text{rev}}^* \in \operatorname{argmax}_{\lambda \geq 0} D_{\text{rev}}(\lambda)$, and the optimal value of the dual function be $D_{\text{rev}}^* := D_{\text{rev}}(\lambda^*)$.

However, it is unclear when $p_{\text{rev}}^*(\cdot; \lambda)$ is a solution to Problem (RU). Due to nonconvexity, $p_{\text{rev}}^*(\cdot; \lambda)$ is not necessarily a solution to Problem (RU), even when we fix $\lambda = \lambda_{\text{rev}}^*$. To address this issue, we next prove that Problem (RU) is strongly dual, despite its nonconvex feasible set.

Theorem 1 (Strong duality). *Let Assumptions 1 and 2 hold. Then, strong duality holds for Problem (RU): $P_{\text{rev}}^* = D_{\text{rev}}^*$. Moreover, an optimal probability measure p^* is given by $p_{\text{rev}}^*(\cdot) = p_{\text{rev}}^*(\cdot; \lambda_{\text{rev}}^*)$.*

We defer the proof of Theorem 1 to Appendix B.1, and sketch our key technical approach below.

A sufficient condition for strong duality to hold for Problem (RU) is that its epigraph is convex (Chamon et al., 2022). To show this, we observe that all probability measures of interest have continuous domains. This allows us to define a non-atomic vector measure in terms of the objective and constraint functions. A key step is a new application of Lyapunov’s convexity theorem (Olech, 1968), which proves the image of a non-atomic vector measure is convex. We stress that it is infeasible to directly apply Lyapunov’s convexity theorem, since the originally-constructed vector measure (Olech, 1968) is not necessarily a probability measure.

Having established strong duality, we evaluate a solution to Problem (RU) given an optimal dual variable λ_{rev}^* in Corollary 1. Denote $\hat{\lambda}_i := \frac{\lambda_i}{1 - \mathbf{1}^\top \lambda}$. When $\mathbf{1}^\top \lambda \neq 1$, we define an

unlearning target $q_u^\dagger(\cdot; \lambda)$ as a ratio of tilted distributions,

$$q_u^\dagger(\cdot; \lambda) := \frac{1}{Z_u^\dagger(\lambda)} \frac{(q(\cdot))^{1/(1 - \mathbf{1}^\top \lambda)}}{\prod_{i=1}^m (q_u^i(\cdot))^{\hat{\lambda}_i}} \quad (2)$$

where $Z_u^\dagger(\lambda) := \int (q(x))^{1/(1 - \mathbf{1}^\top \lambda)} / \prod_{i=1}^m (q_u^i(x))^{\hat{\lambda}_i} dx$. When $\mathbf{1}^\top \lambda = 1$, (2) can be simplified in Appendix B.2; however, this is a degenerate solution, so we omit it here. Since q and $\{q_u^i\}_{i=1}^m$ share the same support, the unlearning target (2) is a proper probability measure. Outside the support, $\frac{0}{0} = 0$ by convention. Viewing q_u^i as a concept distribution to be unlearned, (2) reweights the pretrained model q by the reciprocal of the tilted distribution q_u^i . Hence, (2) excludes, or downweights, samples that are likely in q_u^i . Similar targets have been used in concept erasure (Gandikota et al., 2023) and negation composition (Du et al., 2023).

Corollary 1 (Optimal models). *Let Assumptions 1 and 2 hold. Then, given an optimal dual variable λ_{rev}^* that satisfies $0 \leq \mathbf{1}^\top \lambda_{\text{rev}}^* < 1$, Problem (RU) admits an explicit solution,*

$$p_{\text{rev}}^*(\cdot) = q_u^\dagger(\cdot; \lambda_{\text{rev}}^*). \quad (3)$$

Moreover, it is infeasible or degenerate when $\mathbf{1}^\top \lambda_{\text{rev}}^* \geq 1$.

The proof of Corollary 1 evaluates the partial minimization of $L(p; \lambda)$ over $p \in \Delta$, and we provide it in Appendix B.2. When $0 \leq \mathbf{1}^\top \lambda_{\text{rev}}^* < 1$, the solution (3) captures the effect of the concept unlearning through the distribution ratio (2). An example of this is the concept erasing (Gandikota et al., 2023). We note that when $\mathbf{1}^\top \lambda_{\text{rev}}^* > 1$, the distribution ratio (2) is reversed; this leads to an infeasible unlearning target, which we therefore exclude. Hence, the reverse KL constraints effectively characterize concept unlearning.

2.2. Forward KL-Constrained Unlearning

We next move to Problem (FU). The forward KL objective ensures that the model remains *likely* under the pretrained model, while the m forward KL constraints *reduce* the model’s likelihood of generating samples from m undesirable distributions. To ensure the well-posedness, we assume the following admissibility conditions.

Assumption 3 (Admissibility). *Any probability measure $p \in \Delta$ satisfies two conditions: (i) q and q_u^i are absolutely continuous with respect to p ; and (ii) p is bounded, i.e.,*

$$q, q_u^i \ll p \text{ for } i \in [m] \text{ and } p(x) < \infty \text{ for } x \in X. \quad (4)$$

The absolute continuity ensures the boundedness of forward KL divergence. Problem (FU) is a *non-convex* optimization problem because the superlevel set of forward KL divergence is *nonconvex*. Again, strong duality is not ensured by convex duality analysis (Boyd & Vandenberghe, 2004).

Assumption 4 (Feasibility). *There exists an admissible probability measure $p \in \Delta$: $D_{\text{KL}}(q_u^i \| p) > b_i$ for $i \in [m]$.*

Compared to Assumption 2, we can always find a p such that few of its mass is located on samples with high probability under q_u^i . Thus, Assumption 2 is also practically achievable. Let p_{fw}^* be a solution to Problem (FU) and denote $P_{\text{fw}}^* = D_{\text{KL}}(q \| p_{\text{fw}}^*)$. Denote $\lambda := [\lambda_1, \dots, \lambda_m]^\top$. Let the Lagrangian for Problem (FU) be

$$L_{\text{fw}}(p, \lambda) = D_{\text{KL}}(q \| p) + \sum_{i=1}^m \lambda_i (b_i - D_{\text{KL}}(q_u^i \| p))$$

and the associated dual function is given by $D_{\text{fw}}(\lambda) = \min_{p \in \Delta} L_{\text{fw}}(p, \lambda)$, where the minimization is achieved at $p_{\text{fw}}^*(\cdot; \lambda)$. Let an optimal dual variable be $\lambda_{\text{fw}}^* \in \operatorname{argmax}_{\lambda \geq 0} D_{\text{fw}}(\lambda)$, and the optimal value of the dual function be $D_{\text{fw}}^* := D_{\text{fw}}(\lambda_{\text{fw}}^*)$. Similar to Theorem 1, we next show that $p_{\text{fw}}^*(\cdot; \lambda)$ is a solution to Problem (FU) if $\lambda = \lambda_{\text{fw}}^*$. To do so, we prove that Problem (FU) is strongly dual.

Theorem 2 (Strong duality). *Let Assumptions 3 and 4 hold. Then, strong duality holds for Problem (FU): $P_{\text{fw}}^* = D_{\text{fw}}^*$. Moreover, an optimal probability measure p_{fw}^* is given by $p_{\text{fw}}^*(\cdot) = p_{\text{fw}}^*(\cdot; \lambda_{\text{fw}}^*)$.*

Theorem 2 has a similar proof as the one for Theorem 1. We explain the proof of Theorem 2 in Appendix B.3.

With the strong duality in place, Corollary 2 captures an explicit solution to Problem (FU). When $q(x) - \sum_{i=1}^m \lambda_i q_u^i(x) \geq 0$ for any $x \in X$ and $\lambda \geq 0$, we define an unlearning target $q_u^\triangleleft(\cdot; \lambda)$ as a difference between a pretrained model and a mixture distribution to be unlearned,

$$q_u^\triangleleft(\cdot; \lambda) = \frac{1}{Z_u^\triangleleft(\lambda)} \left(q(\cdot) - \sum_{i=1}^m \lambda_i q_u^i(\cdot) \right) \quad (5)$$

where $Z_u^\triangleleft(\lambda) := \int (q(x) - \sum_{i=1}^m \lambda_i q_u^i(x)) dx$. The unlearning target (5) defines a valid distribution by reducing the likelihood of generating samples from a mixture. This target is similar to the data-unlearning (Wu et al., 2025; Alberti et al., 2025), which runs gradient descent on the retained dataset and gradient ascent on the unlearning dataset.

Corollary 2 (Optimal models). *Let Assumptions 3 and 4 hold. Then, given an optimal dual variable $\lambda_{\text{fw}}^* \geq 0$, Problem (FU) admits an explicit solution,*

$$p_{\text{fw}}^*(\cdot) = q_u^\triangleleft(\cdot; \lambda_{\text{fw}}^*) \quad (6)$$

where λ_{fw}^* satisfies $q(x) - \sum_{i=1}^m \lambda_i q_u^i(x) \geq 0$ for $x \in X$.

We provide the proof of Corollary 2 in Appendix B.4.

2.3. Likelihood-Constrained Unlearning

We now consider Problem (LU). A key distinction from Problem (RU) is that it constrains the likelihood of generating samples from undesirable concepts, rather than the KL

divergences. The likelihood constraints force the model to generate samples from the unlearning concepts less often, with each bounded by ϵ_i . This constraint is intuitive since it prevents the model from assigning high probability to regions associated with the unlearning distributions. Since the likelihood-based feasible set is convex, we can directly conclude strong duality (Boyd & Vandenberghe, 2004).

Assumption 5 (Feasibility). *There exists an admissible probability measure $p \in \Delta$: $\mathbb{E}_p[q_u^i] < \epsilon_i$ for $i \in [m]$.*

Let p_{revl}^* be a solution to Problem (LU) and denote $P_{\text{revl}}^* = D_{\text{KL}}(p_{\text{revl}}^* \| q)$. Denote $\lambda := [\lambda_1, \dots, \lambda_m]^\top$. Let the Lagrangian for Problem (LU) be

$$L_{\text{revl}}(p, \lambda) = D_{\text{KL}}(p \| q) + \sum_{i=1}^m \lambda_i (\mathbb{E}_p[q_u^i] - \epsilon_i)$$

and the associated dual function is given by $D_{\text{revl}}(\lambda) = \min_{p \in \Delta} L_{\text{revl}}(p, \lambda)$, where the minimization is achieved at

$$p_{\text{revl}}^*(\cdot; \lambda) = \frac{1}{Z_u^\circ(\lambda)} q(\cdot) e^{-\sum_{i=1}^m \lambda_i q_u^i(\cdot)} \quad (7)$$

where $Z_u^\circ(\lambda) := \int q(x) e^{-\sum_{i=1}^m \lambda_i q_u^i(x)} dx$. Let an optimal dual variable be $\lambda_{\text{revl}}^* \in \operatorname{argmax}_{\lambda \geq 0} D_{\text{revl}}(\lambda)$, and the optimal value of the dual function be $D_{\text{revl}}^* := D_{\text{revl}}(\lambda_{\text{revl}}^*)$.

Theorem 3 (Strong duality). *Let Assumptions 1 and 5 hold. Then, strong duality holds for Problem (LU): $P_{\text{revl}}^* = D_{\text{revl}}^*$. Moreover, an optimal probability measure p_{revl}^* is given by*

$$p_{\text{revl}}^*(\cdot) = p_{\text{revl}}^*(\cdot; \lambda_{\text{revl}}^*). \quad (8)$$

The proof of Theorem 3 is based on the convex duality analysis (Boyd & Vandenberghe, 2004) in Appendix B.5. Given an optimal dual variable $\lambda_{\text{revl}}^* \geq 0$, the optimal model $p_{\text{revl}}^*(\cdot; \lambda_{\text{revl}}^*)$ that we describe in (7) characterizes the effect of the concept unlearning through the negative exponent $-\sum_{i=1}^m \lambda_i q_u^i$. Regions with high likelihood under the unlearning distributions are downweighted exponentially.

Remark 1 (Constrained vs. Regularized Formulations). *The solutions to Problems (RU), (FU), and (LU) also admit regularized interpretations: each can be viewed as promoting separation from the distributions to be unlearned via suitable penalty terms. Under proper regularity conditions, for each feasible set of constraint thresholds, there exists a set of dual or penalty weights (or exponents) such that the constrained and regularized formulations agree at the level of optimal solutions. However, this equivalence is primarily variational and does not imply that the corresponding methods are equally effective in practice. We highlight three key reasons to prefer constrained formulations.*

- (i) *Constraint thresholds are more interpretable than penalty weights. One can require the learned distribution to be separated from each unlearning distribution*

by a prescribed amount, or put an upper bound on the expected likelihood of undesirable samples. These specifications have direct semantic meaning, whereas penalty weights are indirect and harder to interpret.

- (ii) Penalty weights do not directly specify the desired level of unlearning. A penalty weight, e.g., α_i in the solution $p_{\text{fw}}^*(\cdot)$, does not mean that the i th unlearning target is α_i times more important. Rather, at the optimum, it balances the marginal change in the unlearning objective against the marginal change in the retention objective. In contrast, a constraint level b_i or ϵ_i directly specifies a target separation or likelihood level.
- (iii) Constraints better control multiple unlearning targets. With many concepts or datasets, tuning separate penalty weights is impractical, while assigning the same weight to all targets can lead to uneven unlearning because some concepts are easier to erase than others. Constraint thresholds provide a direct way to enforce balanced unlearning across targets, achieving the desired erasure effect with less deviation from the pretrained model, as shown in Figures 3 and 6.

3. Unlearning for Diffusion Models

We instantiate the three formulations in Section 1.1 by letting all distributions be induced by diffusion models.

3.1. Reverse KL-Constrained Unlearning

We specify the probability measures p , q , and q_u^i in Problem (RU) as the point distributions $p_0(\cdot; s_p)$, $q_0(\cdot; s_q)$, and $q_0^i(\cdot; s_q^i)$, respectively, where s_p , s_q , and s_q^i denote the corresponding score functions. We cast Problem (RU) into a constrained optimization with functional constraints,

$$\underset{s_p \in \mathcal{S}}{\text{minimize}} \quad D_{\text{KL}}(p_0(\cdot; s_p) \parallel q_0(\cdot; s_q)) \quad (9)$$

$$\text{subject to} \quad D_{\text{KL}}(p_0(\cdot; s_p) \parallel q_0^i(\cdot; s_q^i)) \geq b_i \text{ for } i \in [m].$$

Problem (9) is a *nonconvex* optimization problem because the mapping from the score function to the KL divergence is nonlinear. Nevertheless, when viewed as a function of the underlying probability measures, it has the same structure as Problem (RU). To exploit this structure, we assume that the score function class is sufficiently expressive to represent any target point distribution within the class of interest. This is mild, as diffusion models are implemented as overparameterized deep neural networks.

Assumption 6 (Expressivity). *There exists a class of point distributions in which each $p_0(\cdot)$ is induced by a backward process associated with a score function $s_p \in \mathcal{S}$.*

Assumption 7 (Feasibility). *There exists an admissible score s_p : $D_{\text{KL}}(p_0(\cdot; s_p) \parallel q_0^i(\cdot; s_q^i)) > b_i$ for $i \in [m]$.*

Assumption 6 is reasonable, because score-based backward processes converge to a data distribution under very mild conditions (Chen et al., 2023; Li & Yan, 2025). Assumption 7 resembles the rationale behind Assumption 2. Let s_{rev}^* be a solution to Problem (9) and denote $\widehat{P}_{\text{rev}}^* = D_{\text{KL}}(p_0(\cdot; s_{\text{rev}}^*) \parallel q_0(\cdot; s_q))$. The Lagrangian for Problem (9) is given by $\widehat{L}_{\text{rev}}(s_p, \lambda) = L_{\text{rev}}(p_0(\cdot; s_p), \lambda)$, and its dual function is given by $\widehat{D}_{\text{rev}}(\lambda) := \min_{s_p \in \mathcal{S}} L_{\text{rev}}(p_0(\cdot; s_p), \lambda)$. Let an optimal dual variable be $\widehat{\lambda}_{\text{rev}}^* \in \operatorname{argmax}_{\lambda \geq 0} \widehat{D}_{\text{rev}}(\lambda)$, and the optimal value of the dual function be $\widehat{D}_{\text{rev}}^* := \widehat{D}_{\text{rev}}(\widehat{\lambda}_{\text{rev}}^*)$.

Theorem 4 (Strong duality). *Let Assumptions 6 and 7 hold. Then, strong duality holds for Problem (9): $\widehat{P}_{\text{rev}}^* = \widehat{D}_{\text{rev}}^*$.*

We prove Theorem 4 in Appendix C.1. Despite the nonlinearity of Problem (9), we exploit the convexity of a non-atomic vector measure—now it is defined on the path space—to establish strong duality. Since the score function class is sufficiently expressive, we then carry the result over to the score function space using the saddle-point characterization of strong duality. We note that the convexity of a non-atomic vector measure is unaffected by the nonlinearity introduced in (9) when passing to point distributions.

Exploiting strong duality, we introduce a primal-dual algorithm for solving Problem (9), alternating between minimizing the Lagrangian via gradient descent, and maximizing the dual function via dual subgradient ascent:

- (i) Primal step is to solve the Lagrangian problem: $s^+ \in \operatorname{argmin}_s \widehat{L}_{\text{rev}}(s, \lambda)$, where the Lagrangian $\widehat{L}_{\text{rev}}(s, \lambda)$ evaluates the KL divergence to the target (2);
- (ii) Dual step updates the dual variable via $\lambda^+ = \lambda + \eta \partial \widehat{D}_{\text{rev}}(\lambda)$, where the subgradient $\partial \widehat{D}_{\text{rev}}(\lambda)$ is evaluated using the current model s , and η is the stepsize.

See Appendix D for the full description.

3.2. Forward KL-Constrained Unlearning

Instead of the point distributions, we specify the probability measures p , q , q_u^i in Problem (FU) as the path distributions $p_{0:T}(\cdot; s_p)$, $q_{0:T}(\cdot; s_q)$, and $q_{0:T}^i(\cdot; s_q^i)$. Problem (FU) can be cast as a constrained optimization problem,

$$\underset{s_p \in \mathcal{S}}{\text{minimize}} \quad D_{\text{KL}}(q_{0:T}(\cdot; s_q) \parallel p_{0:T}(\cdot; s_p)) \quad (10)$$

$$\text{subject to} \quad D_{\text{KL}}(q_{0:T}^i(\cdot; s_q^i) \parallel p_{0:T}(\cdot; s_p)) \geq b_i \text{ for } i \in [m].$$

Problem (10) is a *nonconvex* optimization problem for the same reason as Problem (9). Similarly, we exploit the strong duality for Problem (FU) to establish strong duality.

Assumption 8 (Expressivity). *There is a class of path distributions in which each $p_{0:T}(\cdot)$ is induced by a backward process associated with a score function $s_p \in \mathcal{S}$.*

Assumption 9 (Feasibility). *There exists an admissible score s_p : $D_{\text{KL}}(q_{0:T}^i(\cdot; s_q^i) \parallel p_{0:T}(\cdot; s_p)) > b_i$ for $i \in [m]$.*

Assumption 8 is more restrictive than Assumption 6, as it uses joint distributions instead of marginal ones. Let s_{fw}^* be a solution to Problem (10) and denote $\hat{P}_{\text{fw}}^* = D_{\text{KL}}(q_{0:T}(\cdot; s_q) \parallel p_{0:T}(\cdot; s_{\text{fw}}^*))$. The Lagrangian for Problem (10) is given by $\hat{L}_{\text{fw}}(s_p, \lambda) = L_{\text{fw}}(p_{0:T}(\cdot; s_p), \lambda)$, and its dual function is given by $\hat{D}_{\text{fw}}(\lambda) := \min_{s_p \in \mathcal{S}} L_{\text{fw}}(p_{0:T}(\cdot; s_p), \lambda)$. Let an optimal dual variable be $\hat{\lambda}_{\text{fw}}^* \in \operatorname{argmax}_{\lambda \geq 0} \hat{D}_{\text{fw}}(\lambda)$, and the optimal value of the dual function be $\hat{D}_{\text{fw}}^* := \hat{D}_{\text{fw}}(\hat{\lambda}_{\text{fw}}^*)$.

Theorem 5 (Strong duality). *Let Assumptions 8 and 9 hold. Then, strong duality holds for Problem (10): $\hat{P}_{\text{fw}}^* = \hat{D}_{\text{fw}}^*$.*

The proof of Theorem 5 in Appendix C.2 is similar to that of Theorem 4, except that the KL divergence is defined over a path distribution. The convexity of a non-atomic vector measure continues to hold. Leveraging strong duality, we introduce a primal-dual algorithm for solving Problem (10):

- (i) Primal step is to solve the Lagrangian problem: $s^+ \in \operatorname{argmin}_s \hat{L}_{\text{fw}}(s, \lambda)$, where the Lagrangian $\hat{L}_{\text{fw}}(s, \lambda)$ has the form of standard score-matching or noise prediction losses;
- (ii) Dual step updates the dual variable via $\lambda^+ = \lambda + \eta \partial \hat{D}_{\text{fw}}(\lambda)$, where the subgradient $\partial \hat{D}_{\text{fw}}(\lambda)$ is evaluated using the current model s , and η is the stepsize.

See Appendix D for the full description.

3.3. Likelihood-Constrained Unlearning

As done in Section 3.1, we introduce the point distributions $q_0^i(\cdot; s_q^i)$ to Problem (LU) to rewrite it as

$$\begin{aligned} & \underset{s_p \in \mathcal{S}}{\text{minimize}} \quad D_{\text{KL}}(p_{0:T}(\cdot; s_p) \parallel q_{0:T}(\cdot; s_q)) & (11) \\ & \text{subject to} \quad \mathbb{E}_{x_0 \sim p_0} [q_0^i(x_0; s_q^i)] \leq \epsilon_i \text{ for } i \in [m]. \end{aligned}$$

In contrast to Problem (LU), Problem (11) is *nonconvex* due to the nonlinear mapping from the score function to the KL divergence and the likelihood. We note that Problems (11) and (9) are similar, differing only in the form of the constraints. Critically, the convexity of a non-atomic vector measure continues to hold despite this change in how the constraints depend on the probability measures.

Assumption 10 (Feasibility). *There exists an admissible score s_p : $\mathbb{E}_{x_0 \sim p_0(\cdot; s_p)} [q_0^i(x_0; s_q^i)] < \epsilon_i$ for $i \in [m]$.*

Let s_{revl}^* be a solution to Problem (11) and denote $\hat{P}_{\text{revl}}^* = D_{\text{KL}}(p_{0:T}(\cdot; s_{\text{revl}}^*) \parallel q_{0:T}(\cdot; s_q))$. The Lagrangian for Problem (11) is given by $\hat{L}_{\text{revl}}(s_p, \lambda) =$

$L_{\text{revl}}(p_{0:T}(\cdot; s_p), \lambda)$, and its dual function is given by $\hat{D}_{\text{revl}}(\lambda) := \min_{s_p \in \mathcal{S}} L_{\text{revl}}(p_{0:T}(\cdot; s_p), \lambda)$. Let an optimal dual variable be $\hat{\lambda}_{\text{revl}}^* \in \operatorname{argmax}_{\lambda \geq 0} \hat{D}_{\text{revl}}(\lambda)$, and the optimal value of the dual function be $\hat{D}_{\text{revl}}^* := \hat{D}_{\text{revl}}(\hat{\lambda}_{\text{revl}}^*)$.

Theorem 6 (Strong duality). *Let Assumptions 6 and 10 hold. Then, strong duality holds for Problem (11): $\hat{P}_{\text{revl}}^* = \hat{D}_{\text{revl}}^*$.*

We prove Theorem 6 in Appendix C.3. The proof is similar to that of Theorem 4, except for the likelihood constraints. Since the likelihood constraints can be lifted to path-space constraints, we can still apply the convexity of a non-atomic vector measure. Because of strong duality, we similarly use a primal-dual algorithm for solving Problem (11):

- (i) Primal step is to solve the Lagrangian problem: $s^+ \in \operatorname{argmin}_s \hat{L}_{\text{revl}}(s, \lambda)$, where the Lagrangian $\hat{L}_{\text{revl}}(s, \lambda)$ involves the computation of the KL divergence and the point probability. To estimate the probability $q_0^i(x_0)$ for a sample x_0 , we apply the information-theoretic diffusion results (Kong et al., 2023) to compute $\log q_0^i(x_0)$ by viewing the pretrained diffusion model as a Gaussian noise channel (see Appendix C.4);
- (ii) Dual step updates the dual variable via $\lambda^+ = \lambda + \eta \partial \hat{D}_{\text{revl}}(\lambda)$, where the subgradient $\partial \hat{D}_{\text{revl}}(\lambda)$ is evaluated using the current model s , and η is the stepsize.

See Appendix D for the full description.

Remark 2 (Parametrization gap). *In Problems (9), (10), and (11), we optimize over distributions induced by diffusion models through a class of score functions. In contrast, the unlearning targets (3), (6), and (8) describe the best unlearning distributions when one can optimize over all admissible probability measures. These targets are attained when the unlearning targets are well represented by the score function parametrization. The remaining discrepancy is a parametrization gap, not a duality gap: strong duality identifies the correct target in the ideal distribution optimization problems, while the diffusion implementation can only approximate these targets.*

In practice, this gap is expected to be small when the desired unlearning target is a smooth perturbation of the pretrained model, the retaining and unlearning distributions are well separated, and the score function class has sufficient expressiveness. It can be large when the constraints are aggressive, retained and unlearned concepts are highly entangled, the target requires sharp density suppression, or the practical parametrization is restrictive, for example because of low-rank adapters, limited optimization, or approximate sampling. Experimentally, a large gap would appear as persistent constraint violation, saturating or oscillating dual variables, a plateau in unlearning performance as thresholds become stricter, or excessive degradation of retention metrics such as KL divergence, KID, or sample quality.

4. Computational Experiments

We demonstrate our constrained unlearning approach through a series of computational experiments using Stable Diffusion v1.4 (Rombach et al., 2022) as the base model. Sections 4.1, 4.2, and 4.3 cover likelihood-constrained concept unlearning, Forward KL-based removal of memorized samples, and reverse KL-based multi-concept unlearning, respectively.

4.1. Likelihood-Constrained Unlearning

We begin with the likelihood-constrained problem (11) by considering a simple setting in which a diffusion model is pretrained on a three-Gaussian mixture, and one component is to be unlearned. As a baseline, we consider concept erasing (Gandikota et al., 2023), which targets the unlearning distribution $(q(\cdot))^{1+\eta}/(q_u(\cdot))^\eta$ for some $\eta > 0$. This is similar to the unlearning target (2) in the reverse KL case. We visualize the distributions of the pretrained and trained models for each approach in the three leftmost plots of Figure 4. At the same unlearning-likelihood level, our likelihood-constrained method achieves a smaller KL divergence to the retained modes, as shown in the right plot of Figure 4, where each point represents a constraint threshold or regularization weight for our method or the baseline.

We further demonstrate this improvement in a pretrained text-to-image model. To unlearn attributes from a concept, we set $q = p_{\text{pre}}(\cdot | c)$, and $q_u^i = p_{\text{pre}}(\cdot | c_i)$, where c is a retained concept (e.g., cowboy) that we aim to stay close to, and c_i denotes other concepts or biases associated with the concept c that we wish to remove (e.g., horse). We show the retention-unlearning tradeoff in the two rightmost plots in Figure 5, where retention performance is measured by the Kernel Inception Distance (KID) (Bińkowski et al., 2021) (Lower KID means better retention performance) and the KL divergence to the pretrained model, respectively. Compared to the baseline, our likelihood-constrained approach achieves strong unlearning (low likelihood) while more effectively limiting deviation from the retained model.

This improved tradeoff can also be qualitatively observed. While both the concept erasure approach and our likelihood constrained approach successfully unlearn the undesired concept, in this case ‘horse’, (Figure 5 Left), our approach keeps the parts of the distribution that we wish to retain, closer to the pre-trained model (Figure 5 Middle left)

4.2. Forward KL-Constrained Unlearning

To demonstrate the effectiveness of the forward KL constraints in (10), we work within the data unlearning paradigm (Alberti et al., 2025), which seeks to prevent the model from generating specific samples that it may have memorized (e.g., for copyright reasons). We use the uncon-

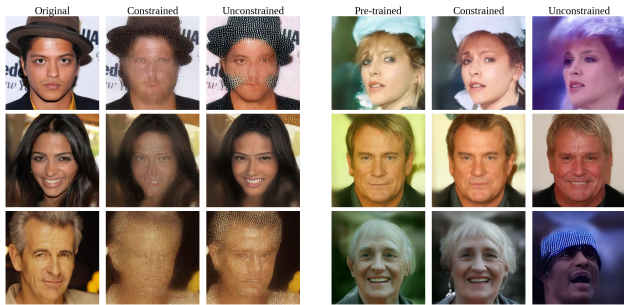


Figure 2. Performance of forward KL-constrained unlearning. (Left) Images generated by pretrained model (three random images to be removed), forward KL-constrained unlearning, and baseline. (Right) Images generated by pretrained model (three random images), forward KL-constrained unlearning, and baseline.

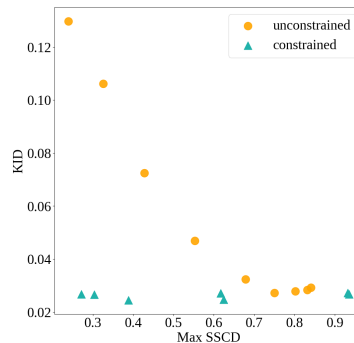


Figure 3. Retention-unlearning tradeoff for forward KL-constrained unlearning and unconstrained baseline. Our constrained model deviates less from the pretrained model at the same level of unlearning (max SSCD).

ditional DDPM trained on the CelebA-HQ dataset as our pretrained model (Ho et al., 2020), and randomly select three images from the dataset to unlearn. As shown in Figure 2, our constrained unlearning approach effectively removes the given samples in the left plot, while retaining the capability of generating other concepts, as shown in the right plot. This significantly improves upon the unconstrained baseline (Alberti et al., 2025), which essentially solves the Lagrangian problem for Problem (10) with equal weights assigned to all of the samples we wish to remove. We also show the retention-unlearning tradeoff in Figure 3, where the unlearning performance is measured by max SSCD (Pizzi et al., 2022) (Higher max SSCD indicates greater similarity to a forgotten sample and therefore weaker unlearning), the retention performance is measured by KID. We see that at the same level of unlearning, measured by max SSCD, our constrained unlearning approach deviates less from the pretrained model, as measured by KID. This is achieved by dynamically assigning large weights to the samples that are hard to unlearn and small weights to the easier samples.

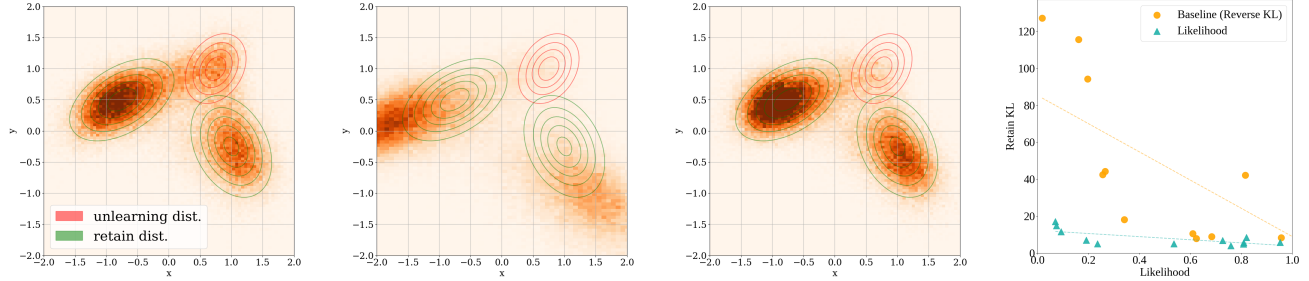


Figure 4. Performance of likelihood-constrained unlearning on a three-Gaussian mixture. From left to right: pretrained model, reverse-KL-constrained unlearning, likelihood-constrained unlearning, and the KL divergence to the retained model versus the likelihood of the unlearning mode.

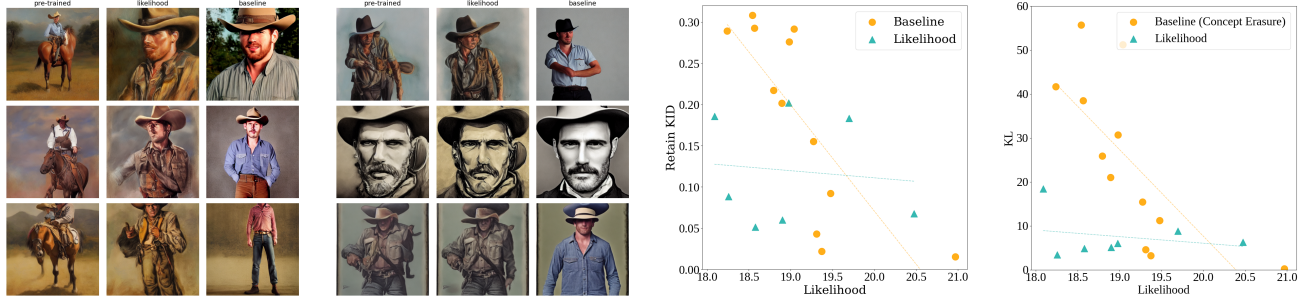


Figure 5. Performance of likelihood-constrained unlearning on a text-to-image model. From left to right: images generated by likelihood-constrained unlearning and baseline for unlearning the concept of ‘horse’ (Left), for retaining the concept of ‘cowboy’ (Middle left), the retention KID (Middle right) and the KL divergence (Right) to the retained model versus the likelihood of unlearning concept.

4.3. Reverse KL-Constrained Unlearning

As in Section 4.1, we consider concept unlearning by setting $q = p_{\text{pre}}(\cdot | c)$, and $q_u^i = p_{\text{pre}}(\cdot | c_i)$ in Problem (9), where c is a retained concept, and c_i denotes other concepts or biases associated with the concept c that we wish to remove. In Figure 6, we show the tradeoff between the worst-case unlearning performance (i.e., minimum KL divergences to unlearning concepts) and the retention performance (i.e., the KL divergence to the pretrained model), with each point representing a constraint threshold or regularization weight. We see that our constrained retention-unlearning tradeoff curve lies below the unconstrained baseline, achieving the same unlearning performance while deviating less from the pretrained model (Figure 6, Left). To confirm this, we also use the maximum CLIP text-to-image similarity score (Hessel et al., 2022) among the unlearning concepts, observing a similar result (Figure 6, Right).

5. Conclusion

We have developed a constrained optimization framework that unifies unlearning as minimizing deviation from a pretrained model, subject to explicit separation constraints from the unlearning distributions. Despite the nonconvexity of these constraints, we prove that constrained unlearning has zero duality gap, which allows us to

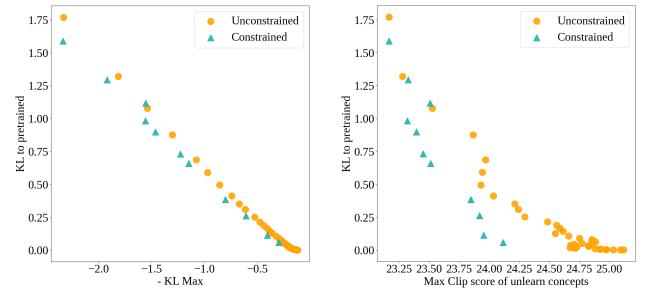


Figure 6. Retention-unlearning tradeoff for reverse KL-constrained unlearning and unconstrained baseline. The KL divergence to the pretrained model versus the KL constraint violation (Left) and the unlearning concept CLIP score (Right).

characterize optimal unlearning targets and develop primal-dual algorithms to approximate them. Empirically, we demonstrate our constrained unlearning approach on image generation tasks, showing a significantly-improved retention-unlearning tradeoff.

Limitations. Despite our theoretical characterizations and promising empirical results, additional experiments are needed to evaluate our method’s effectiveness on unlearning tasks beyond image generation. Further theoretical study is required to establish convergence and sample-complexity guarantees for the proposed primal-dual algorithms.

Impact Statement

This paper aims to improve the responsible behavior of deep generative models by ensuring compliance with unlearning requirements. By establishing a unified optimization framework, our work provides practical guidance for developing more reliable training algorithms, with potential impact across unlearning applications such as language, image, speech, and audio generation.

References

- Alberti, S., Hasanaliyev, K., Shah, M., and Ermon, S. Data unlearning in diffusion models. *arXiv preprint arXiv:2503.01034*, 2025.
- Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. Demystifying mmd gans, 2021. URL <https://arxiv.org/abs/1801.01401>.
- Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.
- Cao, Y. and Yang, J. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.
- Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pp. 5253–5270, 2023.
- Chamon, L. F., Paternain, S., Calvo-Fullana, M., and Ribeiro, A. Constrained learning with non-convex losses. *IEEE Transactions on Information Theory*, 69(3):1739–1760, 2022.
- Chamon, L. F., Karimi, M. R., and Korba, A. Constrained sampling with primal-dual langevin monte carlo. *Advances in Neural Information Processing Systems*, 37:29285–29323, 2024.
- Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023.
- Chen, T., Zhang, S., and Zhou, M. Score forgetting distillation: A swift, data-free method for machine unlearning in diffusion models. *arXiv preprint arXiv:2409.11219*, 2024.
- Christopher, J. K., Baek, S., and Fioretto, N. Constrained synthesis with projected diffusion models. *Advances in Neural Information Processing Systems*, 37:89307–89333, 2024.
- Christopher, J. K., Seamann, A., Cui, J., Khare, S., and Fioretto, F. Constrained diffusion for protein design with hard structural constraints. *arXiv preprint arXiv:2510.14989*, 2025.
- Du, Y., Durkan, C., Strudel, R., Tenenbaum, J. B., Dieleman, S., Fergus, R., Sohl-Dickstein, J., Doucet, A., and Grathwohl, W. S. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and MCMC. In *International conference on machine learning*, pp. 8489–8510. PMLR, 2023.
- European Union. General Data Protection Regulation (GDPR): Regulation (EU) 2016/679. <https://gdpr-info.eu>, 2016.
- Fan, C., Liu, J., Zhang, Y., Wong, E., Wei, D., and Liu, S. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*, 2023a.
- Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., and Lee, K. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models, 2023b. URL <https://arxiv.org/abs/2305.16381>.
- Feng, X., Li, Y., Chen, C., Zhang, L., Li, L., Zhou, J., and Zheng, X. Controllable unlearning for image-to-image generative models via epsilon-constrained optimization. *arXiv preprint arXiv:2408.01689*, 2024.
- Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., and Bau, D. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2426–2436, 2023.
- Gao, H., Pang, T., Du, C., Hu, T., Deng, Z., and Lin, M. Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2131–2141, 2025.
- Golatkar, A., Achille, A., and Soatto, S. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9304–9312, 2020.
- Goldman, E. An introduction to the california consumer privacy act (ccpa). *Santa Clara Univ. Legal Studies Research Paper*, 2020.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015. URL <https://arxiv.org/abs/1512.03385>.
- Heng, A. and Soh, H. Selective amnesia: A continual learning approach to forgetting in deep generative models.

- Advances in Neural Information Processing Systems*, 36: 17170–17194, 2023.
- Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. Clipscore: A reference-free evaluation metric for image captioning, 2022. URL <https://arxiv.org/abs/2104.08718>.
- Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models, 2020. URL <https://arxiv.org/abs/2006.11239>.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021. URL <https://arxiv.org/abs/2106.09685>.
- Khalafi, S., Ding, D., and Ribeiro, A. Constrained diffusion models via dual training. *Advances in Neural Information Processing Systems*, 37:26543–26576, 2024.
- Khalafi, S., Hounie, I., Ding, D., and Ribeiro, A. Composition and alignment of diffusion models using constrained learning. *Advances in Neural Information Processing Systems*, 2025.
- Kong, X., Brekelmans, R., and Steeg, G. V. Information-theoretic diffusion, 2023. URL <https://arxiv.org/abs/2302.03792>.
- Koyejo, O. and Ghosh, J. A representation approach for relative entropy minimization with expectation constraints. In *ICML WDDL workshop*, 2013.
- Lai, C.-H., Song, Y., Kim, D., Mitsufuji, Y., and Ermon, S. The principles of diffusion models. *arXiv preprint arXiv:2510.21890*, 2025.
- Li, B., Gu, R., Wang, J., Qi, L., Li, Y., Wang, R., Qin, Z., and Zhang, T. Towards resilient safety-driven unlearning for diffusion models against downstream fine-tuning. *arXiv preprint arXiv:2507.16302*, 2025.
- Li, G. and Yan, Y. $O(d/t)$ convergence theory for diffusion probabilistic models under minimal assumptions. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Liang, J., Christopher, J. K., Koenig, S., and Fioretto, F. Multi-agent path finding in continuous spaces with projected diffusion models. *arXiv preprint arXiv:2412.17993*, 2024.
- Liang, J., Christopher, J. K., Koenig, S., and Fioretto, F. Simultaneous multi-robot motion planning with projected diffusion models. *arXiv preprint arXiv:2502.03607*, 2025.
- Luan, H., Goh, Y. X., Ng, S.-K., and Ling, C. K. Projected coupled diffusion for test-time constrained joint generation. *arXiv preprint arXiv:2508.10531*, 2025.
- Luo, C. Understanding diffusion models: A unified perspective, 2022. URL <https://arxiv.org/abs/2208.11970>.
- Narasimhan, S. S., Agarwal, S., Rout, L., Shakkottai, S., and Chinchali, S. P. Constrained posterior sampling: Time series generation with hard constraints. *arXiv preprint arXiv:2410.12652*, 2024.
- Olech, C. On the range of an unbounded vector-valued measure. *Mathematical systems theory*, 2(3):251–256, 1968.
- Park, J. and Park, M. Data unlearning beyond uniform forgetting via diffusion time and frequency selection. *arXiv preprint arXiv:2510.17917*, 2025.
- Park, Y.-H., Yun, S., Kim, J.-H., Kim, J., Jang, G., Jeong, Y., Jo, J., and Lee, G. Direct unlearning optimization for robust and safe text-to-image models. *Advances in Neural Information Processing Systems*, 37:80244–80267, 2024.
- Pizzi, E., Roy, S. D., Ravindra, S. N., Goyal, P., and Douze, M. A self-supervised descriptor for image copy detection, 2022. URL <https://arxiv.org/abs/2202.10261>.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2022. URL <https://arxiv.org/abs/2112.10752>.
- Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35: 25278–25294, 2022.
- Wu, J. and Harandi, M. Munba: Machine unlearning via nash bargaining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4754–4765, 2025.
- Wu, J., Le, T., Hayat, M., and Harandi, M. Erasing undesirable influence in diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28263–28273, 2025.
- Zampini, S., Christopher, J. K., Oneto, L., Anguita, D., and Fioretto, F. Training-free constrained generation with stable diffusion models. *arXiv preprint arXiv:2502.05625*, 2025.

Zhang, Y., Chen, X., Jia, J., Zhang, Y., Fan, C., Liu, J., Hong, M., Ding, K., and Liu, S. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *Advances in neural information processing systems*, 37:36748–36776, 2024.

A. Related Work

Unlearning in diffusion models. Our constrained unlearning framework is related to prior work on unlearning in diffusion models. In this setting, given a pretrained diffusion model, unlearning aims to preserve the ability to generate diverse samples while avoiding the generation of samples that resemble specified samples or concepts to be forgotten. To handle these two conflicting objectives, existing approaches primarily adopt a classical regularization strategy, optimizing a weighted sum of retention and forgetting losses (Golatkar et al., 2020; Fan et al., 2023a; Heng & Soh, 2023; Park et al., 2024; Alberti et al., 2025; Wu & Harandi, 2025; Park & Park, 2025; Gao et al., 2025; Li et al., 2025). However, the weighted-sum objective has three key drawbacks: (i) the tradeoff between model utility and unlearning performance is tuned heuristically; (ii) the degree or likelihood of forgetting specific data or concepts is difficult to evaluate; and (iii) multiple unlearning objectives cannot be naturally incorporated within a single regularization term. In contrast, we formulate unlearning from the perspective of constrained distribution optimization. We treat the preservation of pretrained model utility as minimizing the distance between the trained and pretrained models, while expressing unlearning objectives as additional constraints that push the trained model away from specific data or concept distributions. This provides a more principled alternative to existing ad hoc approaches in Gandikota et al. (2023); Alberti et al. (2025). Our unlearning framework (i) offers a theoretical guarantee of an optimal tradeoff between model utility retention and unlearning performance, and (ii) allows for the direct imposition of multiple unlearning constraints. We also remark that our constrained unlearning approach is more tractable, both theoretically and practically, than the bilevel optimization approaches in Zhang et al. (2024); Chen et al. (2024); Wu et al. (2025), which often suffer from high computational costs and limited theoretical guarantees.

Diffusion models under constraints. Our work is also related to recent research on imposing constraints in diffusion models. Two types of constraints have been explored to control the samples generated by diffusion models: (i) hard constraints on individual samples (Liang et al., 2024; 2025; Christopher et al., 2024; Narasimhan et al., 2024; Zampini et al., 2025; Luan et al., 2025; Christopher et al., 2025) and (ii) soft constraints imposed in expectation (Khalafi et al., 2024; Chamon et al., 2024; Khalafi et al., 2025). Our constrained unlearning approach falls into the second category. In contrast to prior work (Khalafi et al., 2024; 2025), our constrained unlearning problem is nonconvex even in the distribution space, which constitutes the main challenge we address. A closely related work is the constrained unlearning formulation of (Feng et al., 2024), assuming a particular class of unlearning distributions. In contrast, our constrained unlearning framework is more general and does not impose such additional restrictions.

B. Proofs in Section 2

B.1. Proof of Theorem 1

Proof. Define $f(p(x)) := p(x) \log \frac{p(x)}{q(x)}$. Since $p, q \ll \mu$, we can rewrite the reverse KL divergence $D_{\text{KL}}(p \parallel q)$ as

$$D_{\text{KL}}(p \parallel q) = \int_X p(x) \log \frac{p(x)}{q(x)} \mu(dx) = \mathbb{E}_\mu [f(p(x))]$$

where $\mu: \mathcal{B}(X) \rightarrow [0, \infty]$ is the standard Lebesgue measure. Similarly, we define $f_i(p(x)) := p(x) \log \frac{p(x)}{q_i^i(x)}$, and thus $D_{\text{KL}}(p \parallel q_u^i) = \mathbb{E}_\mu [f_i(p(x))]$. Meanwhile, we explicitly express the probability constraint for $p \in \Delta$ as

$$\mathbb{E}_\mu [p(x)] := \int_X p(x) \mu(dx) = 1.$$

By relaxing Assumption 1, we define a set of measures \mathcal{P} that contains all probability measures that satisfy (1). To prove strong duality for Problem (RU), we prove it for an equivalent formulation of Problem (RU),

$$\begin{aligned} & \underset{p \in \mathcal{P}}{\text{minimize}} && \mathbb{E}_\mu [f(p(x))] \\ & \text{subject to} && \mathbb{E}_\mu [f_i(p(x))] \geq b_i \quad \text{for } i \in [m] \\ & && \mathbb{E}_\mu [p(x)] = 1. \end{aligned} \tag{12}$$

We note that the key difference between Problem (RU) and Problem (12) is the explicit constraint $\mathbb{E}_\mu [p(x)] = 1$. Thus, they share the optimal solution p_{rev}^* and the optimal value P_{rev}^* . We define the Lagrangian for Problem (12) as $\widehat{L}_{\text{rev}}(p, \lambda, \rho) := L_{\text{rev}}(p, \lambda) + \rho (\mathbb{E}_\mu [p(x)] - 1)$. The associated dual function is $\widehat{D}_{\text{rev}}(\lambda, \rho) = \underset{p \in \mathcal{P}}{\text{minimize}} \widehat{L}_{\text{rev}}(p, \lambda, \rho)$. Denote $(\widehat{\lambda}_{\text{rev}}^*, \widehat{\rho}_{\text{rev}}^*) \in \underset{\lambda \geq 0, \rho}{\text{argmax}} \widehat{D}_{\text{rev}}(\lambda, \rho)$. We can verify that Problem (12) is strongly dual, i.e.,

$$\widehat{D}_{\text{rev}}^* := \widehat{D}_{\text{rev}}(\widehat{\lambda}_{\text{rev}}^*, \widehat{\rho}_{\text{rev}}^*) = P_{\text{rev}}^* := \mathbb{E}_\mu [f(p_{\text{rev}}^*(x))].$$

We provide a strong duality proof in Section B.1.1.

Now, we show that Problem (RU) is strongly dual, i.e., $D_{\text{rev}}^* = \mathbb{E}_\mu [f(p_{\text{rev}}^*(x))]$. We note that weak duality always holds, i.e., $D_{\text{rev}}^* \leq \mathbb{E}_\mu [f(p_{\text{rev}}^*(x))]$. It is sufficient to show $D_{\text{rev}}^* \geq \widehat{D}_{\text{rev}}^*$. By the strong duality for Problem (12), we have

$$\begin{aligned} \widehat{D}_{\text{rev}}^* &= \mathbb{E}_\mu [f(p_{\text{rev}}^*(x))] \\ &= \mathbb{E}_\mu [f(p_{\text{rev}}^*(x))] + \sum_{i=1}^m \widehat{\lambda}_{\text{rev},i}^* (b_i - \mathbb{E}_\mu [f_i(p_{\text{rev}}^*(x))]) \\ &= \underset{p}{\text{minimize}} L_{\text{rev}}(p; \widehat{\lambda}_{\text{rev}}^*) \\ &= D_{\text{rev}}(\widehat{\lambda}_{\text{rev}}^*) \\ &\leq \underset{\lambda \geq 0}{\text{maximize}} D_{\text{rev}}(\lambda) \\ &= D_{\text{rev}}^* \end{aligned}$$

where the 2nd equality is due to that $\mathbb{E}_\mu [p_{\text{rev}}^*(x)] = 1$, the 3rd equality is due to that $(p_{\text{rev}}^*, \widehat{\lambda}_{\text{rev}}^*)$ is a saddle point of $L_{\text{rev}}(p, \lambda)$ for $p \in \Delta$ and $\lambda \geq 0$, the 4th and 5th equalities follow the property of the dual function $D_{\text{rev}}(\lambda)$. Therefore,

$$D_{\text{rev}}^* = \widehat{D}_{\text{rev}}^* = \mathbb{E}_\mu [f(p_{\text{rev}}^*(x))]$$

which proves the strong duality for Problem (RU). \square

B.1.1. STRONG DUALITY FOR PROBLEM (12)

The proof is an application of Lyapunov's convexity theorem in measure theory. We note that weak duality always holds, i.e., $\widehat{D}_{\text{rev}}^* \leq P_{\text{rev}}^*$. Thus, it is sufficient to establish $\widehat{D}_{\text{rev}}^* \geq P_{\text{rev}}^*$. We define an epigraph for Problem (12) as

$$\Gamma := \{(\gamma, t) \mid \exists p \in \mathcal{P}, \mathbb{E}_\mu [f(p(x))] \leq \gamma_0, b_i - \mathbb{E}_\mu [f_i(p(x))] \leq \gamma_i \text{ for } i = 1, \dots, m, \text{ and } \mathbb{E}_\mu [p(x)] - 1 = \tau\}$$

where $\gamma := [\gamma_0, \gamma_1, \dots, \gamma_m]^\top$.

Lemma 1 (Convexity). *Let Assumption 2 hold. Then, the epigraph Γ is non-empty and convex.*

Proof. It is straightforward that Γ is non-empty since Problem (RU) is feasible. We next show that Γ is a convex set. Assume that $(\gamma, \tau), (\gamma', \tau') \in \Gamma$, i.e., there exist $p, p' \in \mathcal{P}$ such that

$$\mathbb{E}_\mu [f(p(x))] \leq \gamma_0, b_i - \mathbb{E}_\mu [f_i(p(x))] \leq \gamma_i \text{ for } i = 1, \dots, m, \text{ and } \mathbb{E}_\mu [p(x)] - 1 = \tau \quad (13a)$$

$$\mathbb{E}_\mu [f(p'(x))] \leq \gamma'_0, b_i - \mathbb{E}_\mu [f_i(p'(x))] \leq \gamma'_i \text{ for } i = 1, \dots, m, \text{ and } \mathbb{E}_\mu [p'(x)] - 1 = \tau'. \quad (13b)$$

It is sufficient to show that $\alpha(\gamma, \tau) + (1 - \alpha)(\gamma', \tau') \in \Gamma$ for $\alpha \in [0, 1]$, i.e., there exists $p_\alpha \in \mathcal{P}$ such that

$$\mathbb{E}_\mu [f(p_\alpha(x))] \leq \alpha\gamma_0 + (1 - \alpha)\gamma'_0 \quad (14a)$$

$$b_i - \mathbb{E}_\mu [f_i(p_\alpha(x))] \leq \alpha\gamma_i + (1 - \alpha)\gamma'_i \text{ for } i = 1, \dots, m \quad (14b)$$

$$\mathbb{E}_\mu [p_\alpha(x)] - 1 = \alpha\tau + (1 - \alpha)\tau'. \quad (14c)$$

To do so, we construct a vector measure $\mathbf{p}: \mathcal{B}(X) \rightarrow \mathbb{R}^{2(m+2)}$,

$$\mathbf{p}(E) = \begin{bmatrix} \int_E f(p(x))\mu(dx) \\ \int_E f_1(p(x))\mu(dx) \\ \vdots \\ \int_E f_m(p(x))\mu(dx) \\ \int_E p(x)\mu(dx) \\ \int_E f(p'(x))\mu(dx) \\ \int_E f_1(p'(x))\mu(dx) \\ \vdots \\ \int_E f_m(p'(x))\mu(dx) \\ \int_E p'(x)\mu(dx) \end{bmatrix}$$

where each entry of \mathbf{p} is a Lebesgue integral. Clearly, $\mathbf{p}(\emptyset) = 0$ and it is non-atomic. When $p \ll q, q_u^i$ (Assumption 1), \mathbf{p} is a proper vector measure, i.e., $\mathbf{p}(X) < \infty$. We note that $\mathbf{p}(X) = \infty$ if and only if p is not absolutely continuous with respect to q or q_u^i , which is an infeasible case of Problem (12). According to a weak version of Lyapunov's convexity theorem (Olech, 1968), there exists a set $E_\alpha \in \mathcal{B}(X)$ such that

$$\mathbf{p}(E_\alpha) = \alpha\mathbf{p}(X) + (1 - \alpha)\mathbf{p}(\emptyset) = \alpha\mathbf{p}(X)$$

for $\alpha \in [0, 1]$. By additivity, $\mathbf{p}(X/E_\alpha) = (1 - \alpha)\mathbf{p}(X)$. Hence, we construct $p_\alpha \in \mathcal{P}$ as follows,

$$p_\alpha(x) = \begin{cases} p(x) & \text{when } x \in E_\alpha \\ p'(x) & \text{when } x \in X/E_\alpha. \end{cases}$$

It is ready to verify that p_α satisfies (14) regarding three conditions as follows.

(i) Condition (14a). By the construction of p_α ,

$$\begin{aligned} \mathbb{E}_\mu [f(p_\alpha(x))] &= \int_{E_\alpha} f(p(x))\mu(dx) + \int_{X/E_\alpha} f(p'(x))\mu(dx) \\ &= \mathbf{p}_1(E_\alpha) + \mathbf{p}_{m+3}(X/E_\alpha) \\ &= \alpha\mathbf{p}_1(X) + (1 - \alpha)\mathbf{p}_{m+3}(X) \\ &= \alpha\mathbb{E}_\mu [f(p(x))] + (1 - \alpha)\mathbb{E}_\mu [f(p'(x))]. \end{aligned}$$

By the assumption (13),

$$\mathbb{E}_\mu [f(p_\alpha(x))] \leq \alpha\gamma_0 + (1 - \alpha)\gamma'_0.$$

(ii) Condition (14b). Similarly, we have

$$\mathbb{E}_\mu [f_i(p_\alpha(x))] = \alpha \mathbb{E}_\mu [f_i(p(x))] + (1 - \alpha) \mathbb{E}_\mu [f_i(p'(x))]$$

By the assumption (13),

$$b_i - \mathbb{E}_\mu [f_i(p_\alpha(x))] \leq \alpha \gamma_i + (1 - \alpha) \gamma'_i.$$

(iii) Condition (14c). Similarly, we have

$$\mathbb{E}_\mu [p_\alpha(x)] - 1 = \alpha \tau + (1 - \alpha) \tau'.$$

Therefore, there exists $p_\alpha \in \mathcal{P}$ such that (14) holds for $\alpha \in [0, 1]$, $(\gamma, \tau), (\gamma', \tau') \in \Gamma$. \square

With Lemma 1, we next show strong duality for Problem (12). First of all, it is easy to check that $(P_{\text{rev}}^*, 0, 0)$ is not an interior point of Γ ; otherwise, there always exists $\epsilon > 0$ such that $(P_{\text{rev}}^* - \epsilon, 0, 0) \in \Gamma$, which contradicts the optimality of P_{rev}^* . By the supporting hyperplane theorem, there exists a non-zero vector $(y, z) \in \mathbb{R}^{m+2}$ such that

$$\sum_{i=0}^m y_i \gamma_i + z \tau \geq y_0 P_{\text{rev}}^* \text{ for } (\gamma, \tau) \in \Gamma. \quad (15)$$

We note that Γ is unbounded above, i.e., for any $(\gamma, \tau) \in \Gamma$, there exists $(\gamma', \tau) \in \Gamma$ such that $\gamma' \geq \gamma$ elementwise. To ensure (15), we must require $y_i \geq 0$ for all i ; otherwise, for some $y_i < 0$, we can set the left-hand side of the inequality (15) to be $-\infty$ by increasing γ_i . Furthermore, we can show that $y_0 > 0$; otherwise, $y_0 = 0$ simplifies (15) to be $\sum_{i=1}^m y_i \gamma_i + z \tau \geq 0$ for $(\gamma, \tau) \in \Gamma$. This contradicts the strict feasibility assumption, since setting (γ, τ) to be the strict feasible point \bar{p} concludes that $\gamma_i < 0$ for all i , and $t = 0$. Hence, $y_0 > 0$, and thus,

$$\gamma_0 + \sum_{i=1}^m \bar{y}_i \gamma_i + \bar{z} \tau \geq P_{\text{rev}}^* \text{ for } (\gamma, \tau) \in \Gamma$$

where $\bar{y}_i := \frac{y_i}{y_0}$ and $\bar{z} := \frac{z}{y_0}$. It further implies that

$$\mathbb{E}_\mu [f(p(x))] + \sum_{i=1}^m \bar{y}_i (b_i - \mathbb{E}_\mu [f_i(p(x))]) + \bar{z} (\mathbb{E}_\mu [p(x)] - 1) \geq P_{\text{rev}}^*$$

or, equivalently, $\widehat{L}_{\text{rev}}(p; \bar{y}, \bar{z}) \geq P_{\text{rev}}^*$ for $p \in \mathcal{P}$ and $\bar{y} \geq 0$. Thus, minimization of $\widehat{L}_{\text{rev}}(p; \bar{y}, \bar{z})$ over $p \in \mathcal{P}$ leads to $\widehat{D}_{\text{rev}}(\bar{y}, \bar{z}) \geq P_{\text{rev}}^*$, and thus $\widehat{D}_{\text{rev}}^* \geq P_{\text{rev}}^*$.

B.2. Proof of Corollary 1

Proof. When $1 - \mathbf{1}^\top \lambda \neq 0$, we rearrange the Lagrangian $L_{\text{rev}}(p, \lambda)$ as follows,

$$\begin{aligned} L_{\text{rev}}(p, \lambda) &= D_{\text{KL}}(p \| q) + \sum_{i=1}^m \lambda_i (b_i - D_{\text{KL}}(p \| q_u^i)) \\ &= \mathbb{E}_{x \sim p} [\log p(x) - \log q(x)] + \lambda^\top b - \mathbf{1}^\top \lambda \mathbb{E}_{x \sim p} [\log p(x)] + \sum_{i=1}^m \lambda_i \mathbb{E}_{x \sim p} [\log q_u^i(x)] \\ &= \lambda^\top b + (1 - \mathbf{1}^\top \lambda) \mathbb{E}_{x \sim p} [\log p(x)] - \left(\mathbb{E}_{x \sim p} [\log q(x)] - \sum_{i=1}^m \lambda_i \mathbb{E}_{x \sim p} [\log q_u^i(x)] \right) \\ &= \lambda^\top b + (1 - \mathbf{1}^\top \lambda) \left(\mathbb{E}_{x \sim p} [\log p(x)] - \mathbb{E}_{x \sim p} \left[\log \frac{(q(x))^{1/(1-\mathbf{1}^\top \lambda)}}{\prod_{i=1}^m (q_u^i(x))^{\lambda_i}} \right] \right) \\ &= \lambda^\top b + (1 - \mathbf{1}^\top \lambda) (D_{\text{KL}}(p(\cdot) \| q_u^\dagger(\cdot; \lambda)) - \log Z_u^\dagger(\lambda)) \end{aligned}$$

where the 4th equality abbreviates $\widehat{\lambda}_i := \frac{\lambda_i}{1 - \mathbf{1}^\top \lambda}$ for $i = 1, \dots, m$, and the last equality is due to that q_u^\dagger is well-defined for q and q_u^i that share the same support. When $1 - \mathbf{1}^\top \lambda = 0$, it is straightforward to compute

$$\begin{aligned} L_{\text{rev}}(p, \lambda) &= \lambda^\top b - \mathbb{E}_{x \sim p} \left[\log \frac{q(x)}{\prod_{i=1}^m (q_u^i(x))^{\lambda_i}} \right] \\ &= \mathbf{1}^\top b - \mathbb{E}_{x \sim p} [\log q_u^\dagger(x; \lambda)] - \log Z_u^\dagger(\lambda) \end{aligned}$$

and simplify the unlearning target q_u^\dagger as

$$q_u^\dagger(\cdot; \lambda) := \frac{1}{Z_u^\dagger(\lambda)} \frac{q(\cdot)}{\prod_{i=1}^m (q_u^i(\cdot))^{\lambda_i}}$$

where $Z_u^\dagger(\lambda) := \int q(x) / \prod_{i=1}^m (q_u^i(x))^{\lambda_i} dx$.

To evaluate the dual function $D_{\text{rev}}(\lambda) := \text{minimize}_{p \in \Delta} L_{\text{rev}}(p, \lambda)$, there are three cases.

(i) When $0 \leq \mathbf{1}^\top \lambda < 1$, minimization of $L_{\text{rev}}(p, \lambda)$ over p is at a unique minimizer,

$$p_{\text{rev}}^*(\cdot; \lambda) = q_u^\dagger(\cdot; \lambda)$$

which defines the dual function $D_{\text{rev}}(\lambda) = \lambda^\top b - (1 - \mathbf{1}^\top \lambda) \log Z_u^\dagger(\lambda)$.

(ii) When $\mathbf{1}^\top \lambda = 1$, minimization of $L_{\text{rev}}(p, \lambda)$ over p is at $X(\lambda) := \text{argmax}_x q(x) / \prod_{i=1}^m (q_u^i(x))^{\lambda_i}$. Thus,

$$p_{\text{rev}}^*(\cdot; \lambda) = \frac{1}{q_u^\dagger(X(\lambda); \lambda)} \delta_{X(\lambda)}(\cdot).$$

Hence,

$$D_{\text{rev}}(\lambda) = \mathbf{1}^\top b - \frac{1}{q_u^\dagger(X(\lambda); \lambda)} \int_{X(\lambda)} \log q_u^\dagger(x; \lambda) dx - \log Z_u^\dagger(\lambda).$$

(iii) When $\mathbf{1}^\top \lambda > 1$, minimization of $L_{\text{rev}}(p, \lambda)$ over p is equivalent to

$$\text{maximize}_{p \in \Delta} D_{\text{KL}}(p \| q_u^\dagger).$$

We note that when $p \in \Delta$ does not satisfy Assumption 1, $D_{\text{KL}}(p \| q_u^\dagger) = \infty$. Hence, $D_{\text{rev}}^* = -\infty$, which contradicts to Assumption 2.

Hence, we can summarize them as follows. When $\mathbf{1}^\top \lambda \leq 1$ and $\lambda \geq 0$, the dual function $D_{\text{rev}}(\lambda)$ has an explicit form,

$$D_{\text{rev}}(\lambda) = \begin{cases} \lambda^\top b - (1 - \mathbf{1}^\top \lambda) \log Z_u^\dagger(\lambda) & \text{when } 0 \leq \mathbf{1}^\top \lambda < 1 \text{ and } \lambda \geq 0 \\ \mathbf{1}^\top b - \frac{1}{q_u^\dagger(X(\lambda); \lambda)} \int_{X(\lambda)} \log q_u^\dagger(x; \lambda) dx - \log Z_u^\dagger(\lambda) & \text{when } \mathbf{1}^\top \lambda = 1 \text{ and } \lambda \geq 0 \end{cases}$$

where $X(\lambda) := \text{argmax}_x q(x) / \prod_{i=1}^m (q_u^i(x))^{\lambda_i}$. The Lagrangian function $L_{\text{rev}}(p; \lambda)$ has an explicit form,

$$L_{\text{rev}}(p, \lambda) = \begin{cases} \lambda^\top b + (1 - \mathbf{1}^\top \lambda) \left(D_{\text{KL}}(p \| q_u^\dagger) - \log Z_u^\dagger(\lambda) \right) & \text{when } \mathbf{1}^\top \lambda \neq 1 \text{ and } \lambda \geq 0 \\ \mathbf{1}^\top b - \mathbb{E}_{x \sim p} \left[\log q_u^\dagger(x; \lambda) \right] - \log Z_u^\dagger(\lambda) & \text{when } \mathbf{1}^\top \lambda = 1 \text{ and } \lambda \geq 0. \end{cases}$$

A partial minimizer of the Lagrangian $L_{\text{rev}}(p; \lambda)$ over p is achieved at

$$p_{\text{rev}}^*(\cdot; \lambda) = \begin{cases} q_u^\dagger(\cdot; \lambda) & \text{when } 0 \leq \mathbf{1}^\top \lambda < 1 \text{ and } \lambda \geq 0 \\ \frac{1}{q_u^\dagger(X(\lambda); \lambda)} \delta_{X(\lambda)}(\cdot) & \text{when } \mathbf{1}^\top \lambda = 1 \text{ and } \lambda \geq 0. \end{cases}$$

Finally, we apply the strong duality in Theorem 1 to evaluate $p_{\text{rev}}^*(\cdot) = p_{\text{rev}}^*(\cdot; \lambda_{\text{rev}}^*)$. When $\mathbf{1}^\top \lambda_{\text{rev}}^* = 1$, the solution p_{rev}^* is a delta function, which is not practically attainable. This case can always be avoided by scaling λ_{rev}^* , allowing us to focus on $0 \leq \mathbf{1}^\top \lambda_{\text{rev}}^* < 1$. \square

B.3. Proof of Theorem 2

Proof. Since the proof is similar to the proof of Theorem 1, we omit the detail and only state the key changes. Define $f(p(x)) := q(x) \log \frac{q(x)}{p(x)}$. Since $p, q \ll \mu$, we can rewrite the forward KL divergence $D_{\text{KL}}(q \| p)$ as

$$D_{\text{KL}}(q \| p) = \int_X q(x) \log \frac{q(x)}{p(x)} \mu(dx) = \mathbb{E}_\mu [f(p(x))]$$

where $\mu: \mathcal{B}(X) \rightarrow [0, \infty]$ is the standard Lebesgue measure. Similarly, we define $f_i(p(x)) := q_u^i(x) \log \frac{q_u^i(x)}{p(x)}$, and thus $D_{\text{KL}}(q_u^i \| p) = \mathbb{E}_\mu [f_i(p(x))]$. Meanwhile, we explicitly express the probability constraint for $p \in \Delta_X$ as

$$\mathbb{E}_\mu [p(x)] := \int_X p(x) \mu(dx) = 1.$$

With a slight abuse of notation, we also denote the set of all measures that satisfy (4) by \mathcal{P} . To prove strong duality for Problem (FU), we prove it for an equivalent formulation of Problem (FU),

$$\begin{aligned} & \underset{p \in \mathcal{P}}{\text{minimize}} && \mathbb{E}_\mu [f(p(x))] \\ & \text{subject to} && \mathbb{E}_\mu [f_i(p(x))] \geq b_i \quad \text{for } i \in [m] \\ & && \mathbb{E}_\mu [p(x)] = 1. \end{aligned} \tag{16}$$

The rest follows the same proof steps in the proof of Theorem 1. First, we employ Lyapunov's convexity theorem to show that the epigraph for Problem (16) is non-empty and convex, implying that (16) is strongly dual. Second, we translate the strong duality for Problem (16) to the original problem (FU). \square

B.4. Proof of Corollary 2

Proof. For any $\lambda \geq 0$, we rearrange the Lagrangian $L_{\text{fw}}(p, \lambda)$ as follows,

$$\begin{aligned} L_{\text{fw}}(p, \lambda) &= D_{\text{KL}}(q \| p) + \sum_{i=1}^m \lambda_i (b_i - D_{\text{KL}}(q_u^i \| p)) \\ &= \mathbb{E}_{x \sim q} [\log q(x)] + \sum_{i=1}^m \lambda_i (b_i - \mathbb{E}_{x \sim q_u^i} [\log q_u^i(x)]) - \int \left(q(x) - \sum_{i=1}^m \lambda_i q_u^i(x) \right) \log p(x) \\ &= \mathbb{E}_{x \sim q} [\log q(x)] + \sum_{i=1}^m \lambda_i (b_i - \mathbb{E}_{x \sim q_u^i} [\log q_u^i(x)]) - Z_u^\triangleleft(\lambda) \int q_u^\triangleleft(x; \lambda) \log p(x) \\ &= \mathbb{E}_{x \sim q} [\log q(x)] + \sum_{i=1}^m \lambda_i (b_i - \mathbb{E}_{x \sim q_u^i} [\log q_u^i(x)]) - Z_u^\triangleleft(\lambda) \mathbb{E}_{x \sim q_u^\triangleleft(\cdot; \lambda)} [\log q_u^\triangleleft(x; \lambda)] \\ &\quad + Z_u^\triangleleft(\lambda) D_{\text{KL}}(q_u^\triangleleft(\cdot; \lambda) \| p(\cdot)). \end{aligned}$$

It is important to verify that $q(x) - \sum_{i=1}^m \lambda_{\text{fw},i}^* q_u^i(x) \geq 0$ for any $x \in X$. If not, then $q(x) - \sum_{i=1}^m \lambda_{\text{fw},i}^* q_u^i(x) < 0$ for some $x \in X$. In this case, minimization of $L_{\text{fw}}(p, \lambda_{\text{fw}}^*)$ is achieved by p_{fw}^* that satisfies $p_{\text{fw}}^*(x) = 0$, and $L_{\text{fw}}(p_{\text{fw}}^*, \lambda_{\text{fw}}^*) = -\infty$. However, from the strong duality we have $L_{\text{fw}}(p_{\text{fw}}^*, \lambda_{\text{fw}}^*) = D_{\text{KL}}(q \| p_{\text{fw}}^*) \geq 0$, leading to a contradiction. Therefore, we apply the strong duality from Theorem 2 to evaluate,

$$p_{\text{fw}}^*(\cdot) = p_{\text{fw}}^*(\cdot; \lambda_{\text{fw}}^*) = \underset{p \in \Delta}{\text{argmin}} L_{\text{fw}}(p, \lambda_{\text{fw}}^*) = q_u^\triangleleft(\cdot; \lambda_{\text{fw}}^*)$$

which completes the proof. \square

B.5. Proof of Theorem 3

Proof. We note that Problem (LU) is a convex optimization problem, since the reverse KL divergence is a strongly convex function and the likelihood constraints are linear. According to the standard duality analysis (Boyd & Vandenberghe, 2004), there is no duality gap when the problem is strictly feasible. \square

C. Proofs in Section 3

C.1. Proof of Theorem 4

Proof. We consider Problem (9) in the distribution space,

$$\begin{aligned} & \underset{p_0(\cdot)}{\text{minimize}} && D_{\text{KL}}(p_0(\cdot) \parallel q_0(\cdot; s_q)) \\ & \text{subject to} && D_{\text{KL}}(p_0(\cdot) \parallel q_0^i(\cdot; s_q^i)) \geq b_i \text{ for } i \in [m]. \end{aligned} \quad (17)$$

Define $f(p_{0:T}(x_{0:T})) := p_{0:T}(x_{0:T}) \log \frac{\int p_{0:T}(x_{0:T}) dx_{1:T}}{\int q_{0:T}(x_{0:T}) dx_{1:T}}$. Since $p_{0:T}, q_{0:T} \ll \mu$, we can rewrite the reverse KL divergence $D_{\text{KL}}(p_0 \parallel q_0)$ as

$$D_{\text{KL}}(p_0 \parallel q_0) = \int_{X^{T+1}} p_{0:T}(x_{0:T}) \log \frac{\int p_{0:T}(x_{0:T}) dx_{1:T}}{\int q_{0:T}(x_{0:T}) dx_{1:T}} \mu(dx_{0:T}) = \mathbb{E}_\mu [f(p_{0:T}(x_{0:T}))]$$

where $\mu: \mathcal{B}(X^{T+1}) \rightarrow [0, \infty]$ is the standard Lebesgue measure. Similarly, we define $f_i(p_{0:T}(x_{0:T})) := p_{0:T}(x_{0:T}) \log \frac{\int p_{0:T}(x_{0:T}) dx_{1:T}}{\int q_{0:T}^i(x_{0:T}) dx_{1:T}}$, and thus $D_{\text{KL}}(p_0 \parallel q_0^i) = \mathbb{E}_\mu [f_i(p_{0:T}(x_{0:T}))]$. Meanwhile, we explicitly express the probability constraint for $p_{0:T} \in \Delta(X^{T+1})$ as

$$\mathbb{E}_\mu [p_{0:T}(x_{0:T})] := \int_{X^{T+1}} p_{0:T}(x_{0:T}) \mu(dx_{0:T}) = 1.$$

To prove strong duality for Problem (17), we prove it for an equivalent formulation of Problem (17),

$$\begin{aligned} & \underset{p_{0:T} \in \mathcal{P}}{\text{minimize}} && \mathbb{E}_\mu [f(p_{0:T}(x_{0:T}))] \\ & \text{subject to} && \mathbb{E}_\mu [f_i(p_{0:T}(x_{0:T}))] \geq b_i \text{ for } i \in [m] \\ & && \mathbb{E}_\mu [p_{0:T}(x_{0:T})] = 1. \end{aligned} \quad (18)$$

The rest follows the same proof steps in the proof of Theorem 1. First, we employ Lyapunov's convexity theorem to show that the epigraph for Problem (18) is non-empty and convex, implying that (18) is strongly dual. Second, we translate the strong duality for Problem (18) to the original problem (17).

To prove strong duality for Problem (9), we introduce some notation as follows. Let $\bar{p}_{0,\text{rev}}^*$ be a solution to Problem (17) and denote $\bar{P}_{\text{rev}}^* := D_{\text{KL}}(\bar{p}_{0,\text{rev}}^*(\cdot) \parallel q_0(\cdot; s_q))$. The Lagrangian for Problem (17) is given by $\bar{L}_{\text{rev}}(p_0(\cdot), \lambda) := L_{\text{rev}}(p_0(\cdot), \lambda)$, and its dual function is given by $\bar{D}_{\text{rev}}(\lambda) := \min_{p_0} \bar{L}_{\text{rev}}(p_0(\cdot), \lambda)$. Let an optimal dual variable be $\bar{\lambda}_{\text{rev}}^* \in \arg\max_{\lambda \geq 0} \bar{D}_{\text{rev}}(\lambda)$, and the optimal value of the dual function be $\bar{D}_{\text{rev}}^* := \bar{D}_{\text{rev}}(\bar{\lambda}_{\text{rev}}^*)$. From the strong duality of Problem (17), there exists a pair $(\bar{p}_{0,\text{rev}}^*, \bar{\lambda}_{\text{rev}}^*)$ such that

$$\bar{P}_{\text{rev}}^* = \bar{D}_{\text{rev}}^* \text{ or } D_{\text{KL}}(\bar{p}_{0,\text{rev}}^*(\cdot) \parallel q_0(\cdot; s_q)) = \bar{D}_{\text{rev}}(\bar{\lambda}_{\text{rev}}^*).$$

With Assumption 7, this implies that $(\bar{p}_{0,\text{rev}}^*, \bar{\lambda}_{\text{rev}}^*)$ is a saddle point of the Lagrangian $\bar{L}_{\text{rev}}(p_0(\cdot), \lambda)$,

$$\bar{L}_{\text{rev}}(\bar{p}_{0,\text{rev}}^*(\cdot), \lambda) \leq \bar{L}_{\text{rev}}(\bar{p}_{0,\text{rev}}^*(\cdot), \bar{\lambda}_{\text{rev}}^*) \leq \bar{L}_{\text{rev}}(p_0(\cdot), \bar{\lambda}_{\text{rev}}^*) \text{ for all } p_0(\cdot) \text{ and } \lambda \geq 0.$$

By Assumption 6, the score function class \mathcal{S} is expressive enough, any point distribution $p_0(\cdot)$ can be represented as $p_0(\cdot; s_p)$ with some $s_p \in \mathcal{S}$; and vice versa. Thus, we can express $\bar{p}_{0,\text{rev}}^*(\cdot)$ as $p_0(\cdot; s_{\text{rev}}^*)$ with some $s_{\text{rev}}^* \in \mathcal{S}$. We also note that the dual function $\bar{D}_{\text{rev}}(\lambda)$ in the path and score function spaces are the same. Hence, the optimal dual function for Problem (9) remains to be $\widehat{D}_{\text{rev}}(\lambda) = \bar{D}_{\text{rev}}(\lambda)$. Thus, $(s_{\text{rev}}^*, \bar{\lambda}_{\text{rev}}^*)$ is a saddle point of the Lagrangian $\widehat{L}_{\text{rev}}(s_p; \lambda) := \bar{L}_{\text{rev}}(p_0(\cdot; s_p); \lambda)$,

$$\widehat{L}_{\text{rev}}(s_{\text{rev}}^*; \lambda) \leq \widehat{L}_{\text{rev}}(s_{\text{rev}}^*; \bar{\lambda}_{\text{rev}}^*) \leq \widehat{L}_{\text{rev}}(s_p; \bar{\lambda}_{\text{rev}}^*) \text{ for all } s_p \in \mathcal{S} \text{ and } \lambda \geq 0.$$

Therefore, the strong duality holds for Problem (9) in the score function space. \square

C.2. Proof of Theorem 5

Proof. We consider Problem (10) in the path distribution space,

$$\begin{aligned} & \underset{p_{0:T}(\cdot)}{\text{minimize}} && D_{\text{KL}}(q_{0:T}(\cdot; s_q) \parallel p_{0:T}(\cdot)) \\ & \text{subject to} && D_{\text{KL}}(q_{0:T}^i(\cdot; s_q^i) \parallel p_{0:T}(\cdot)) \geq b_i \text{ for } i \in [m]. \end{aligned} \quad (19)$$

Define $f(p_{0:T}(x_{0:T})) := q_{0:T}(x_{0:T}) \log \frac{q_{0:T}(x_{0:T})}{p_{0:T}(x_{0:T})}$. Since $p_{0:T}, q_{0:T} \ll \mu$, we can rewrite the forward KL divergence $D_{\text{KL}}(q_{0:T} \parallel p_{0:T})$ as

$$D_{\text{KL}}(q_{0:T} \parallel p_{0:T}) = \int_{X^{T+1}} q_{0:T}(x_{0:T}) \log \frac{q_{0:T}(x_{0:T})}{p_{0:T}(x_{0:T})} \mu(dx_{0:T}) = \mathbb{E}_\mu [f(p_{0:T}(x_{0:T}))]$$

where $\mu: \mathcal{B}(X^{T+1}) \rightarrow [0, \infty]$ is the standard Lebesgue measure. Similarly, we define $f_i(p_{0:T}(x_{0:T})) := q_{0:T}^i(x_{0:T}) \log \frac{q_{0:T}^i(x_{0:T})}{p_{0:T}(x_{0:T})}$, and thus $D_{\text{KL}}(q_{0:T}^i \parallel p_{0:T}) = \mathbb{E}_\mu [f_i(p_{0:T}(x_{0:T}))]$. Meanwhile, we explicitly express the probability constraint for $p_{0:T} \in \Delta(X^{T+1})$ as

$$\mathbb{E}_\mu [p_{0:T}(x_{0:T})] := \int_{X^{T+1}} p_{0:T}(x_{0:T}) \mu(dx_{0:T}) = 1.$$

To prove strong duality for Problem (19), we prove it for an equivalent formulation of Problem (19),

$$\begin{aligned} & \underset{p_{0:T} \in \mathcal{P}}{\text{minimize}} && \mathbb{E}_\mu [f(p_{0:T}(x_{0:T}))] \\ & \text{subject to} && \mathbb{E}_\mu [f_i(p_{0:T}(x_{0:T}))] \geq b_i \text{ for } i \in [m] \\ & && \mathbb{E}_\mu [p_{0:T}(x_{0:T})] = 1. \end{aligned} \quad (20)$$

The rest follows the same proof steps in the proof of Theorem 2. First, we employ Lyapunov's convexity theorem to show that the epigraph for Problem (20) is non-empty and convex, implying that (20) is strongly dual. Second, we translate the strong duality for Problem (20) to the original problem (19).

To prove strong duality for Problem (20), we introduce some notation as follows. Let $\bar{p}_{0:T,\text{fw}}^*$ be a solution to Problem (20) and denote $\bar{P}_{\text{fw}}^* := D_{\text{KL}}(q_{0:T}(\cdot; s_q) \parallel \bar{p}_{0:T,\text{fw}}^*(\cdot))$. The Lagrangian for Problem (19) is given by $\bar{L}_{\text{fw}}(p_{0:T}(\cdot), \lambda) := L_{\text{fw}}(p_{0:T}(\cdot), \lambda)$, and its dual function is given by $\bar{D}_{\text{fw}}(\lambda) := \min_{p_{0:T}} \bar{L}_{\text{fw}}(p_{0:T}(\cdot), \lambda)$. Let an optimal dual variable be $\bar{\lambda}_{\text{fw}}^* \in \operatorname{argmax}_{\lambda \geq 0} \bar{D}_{\text{fw}}(\lambda)$, and the optimal value of the dual function be $\bar{D}_{\text{fw}}^* := \bar{D}_{\text{fw}}(\bar{\lambda}_{\text{fw}}^*)$. From the strong duality of Problem (20), there exists a pair $(\bar{p}_{0:T,\text{fw}}^*, \bar{\lambda}_{\text{fw}}^*)$ such that

$$\bar{P}_{\text{fw}}^* = \bar{D}_{\text{fw}}^* \text{ or } D_{\text{KL}}(q_{0:T}(\cdot; s_q) \parallel \bar{p}_{0:T,\text{fw}}^*(\cdot)) = \bar{D}_{\text{fw}}(\bar{\lambda}_{\text{fw}}^*).$$

With Assumption 9, this implies that $(\bar{p}_{0:T,\text{fw}}^*, \bar{\lambda}_{\text{fw}}^*)$ is a saddle point of the Lagrangian $\bar{L}_{\text{fw}}(p_{0:T}(\cdot), \lambda)$,

$$\bar{L}_{\text{fw}}(\bar{p}_{0:T,\text{fw}}^*(\cdot), \lambda) \leq \bar{L}_{\text{fw}}(\bar{p}_{0:T,\text{fw}}^*(\cdot), \bar{\lambda}_{\text{fw}}^*) \leq \bar{L}_{\text{fw}}(p_{0:T}(\cdot), \bar{\lambda}_{\text{fw}}^*) \text{ for all } p_{0:T}(\cdot) \text{ and } \lambda \geq 0.$$

By Assumption 8, the score function class \mathcal{S} is expressive enough, any path distribution $p_{0:T}(\cdot)$ can be represented as $p_{0:T}(\cdot; s_p)$ with some $s_p \in \mathcal{S}$; and vice versa. Thus, we can express $\bar{p}_{0:T,\text{fw}}^*(\cdot)$ as $p_{0:T}(\cdot; s_{\text{fw}}^*)$ with some $s_{\text{fw}}^* \in \mathcal{S}$. We also note that the dual function $\bar{D}_{\text{fw}}(\lambda)$ in the path and score function spaces are the same. Hence, the optimal dual function for Problem (10) remains to be $\hat{D}_{\text{fw}}(\lambda) = \bar{D}_{\text{fw}}(\lambda)$. Thus, $(s_{\text{fw}}^*, \bar{\lambda}_{\text{fw}}^*)$ is a saddle point of the Lagrangian $\hat{L}_{\text{fw}}(s_p; \lambda) := \bar{L}_{\text{fw}}(p_{0:T}(x_{0:T}; s_p); \lambda)$,

$$\hat{L}_{\text{fw}}(s_{\text{fw}}^*; \lambda) \leq \hat{L}_{\text{fw}}(s_{\text{fw}}^*; \bar{\lambda}_{\text{fw}}^*) \leq \hat{L}_{\text{fw}}(s_p; \bar{\lambda}_{\text{fw}}^*) \text{ for all } s_p \in \mathcal{S} \text{ and } \lambda \geq 0.$$

Therefore, the strong duality holds for Problem (10) in the score function space. \square

C.3. Proof of Theorem 6

Proof. We consider Problem (11) in the distribution space,

$$\begin{aligned} & \underset{p_{0:T}(\cdot)}{\text{minimize}} && D_{\text{KL}}(p_{0:T}(\cdot) \parallel q_{0:T}(\cdot; s_q)) \\ & \text{subject to} && \mathbb{E}_{x_0 \sim p_0} [q_0^i(x_0; s_q^i)] \leq \epsilon_i \text{ for } i \in [m]. \end{aligned} \quad (21)$$

Define $f(p_{0:T}(x_{0:T})) := p_{0:T}(x_{0:T}) \log \frac{p_{0:T}(x_{0:T})}{q_{0:T}(x_{0:T})}$. Since $p_{0:T}, q_{0:T} \ll \mu$, we can rewrite the reverse KL divergence $D_{\text{KL}}(p_0 \parallel q_0)$ as

$$D_{\text{KL}}(p_{0:T} \parallel q_{0:T}) = \int_{X^{T+1}} p_{0:T}(x_{0:T}) \log \frac{p_{0:T}(x_{0:T})}{q_{0:T}(x_{0:T})} \mu(dx_{0:T}) = \mathbb{E}_\mu [f(p_{0:T}(x_{0:T}))]$$

where $\mu: \mathcal{B}(X^{T+1}) \rightarrow [0, \infty]$ is the standard Lebesgue measure. Similarly, we define $f_i(p_{0:T}(x_{0:T})) := p_{0:T}(x_{0:T}) \int q_{0:T}^i(x_{0:T}) dx_{1:T}$, and thus $\mathbb{E}_{x_0 \sim p_0} [q_0^i(x_0; s_q^i)] = \mathbb{E}_\mu [f_i(p_{0:T}(x_{0:T}))]$. Meanwhile, we explicitly express the probability constraint for $p_{0:T} \in \Delta(X^{T+1})$ as

$$\mathbb{E}_\mu [p_{0:T}(x_{0:T})] := \int_{X^{T+1}} p_{0:T}(x_{0:T}) \mu(dx_{0:T}) = 1.$$

To prove strong duality for Problem (21), we prove it for an equivalent formulation of Problem (21),

$$\begin{aligned} & \underset{p_{0:T} \in \mathcal{P}}{\text{minimize}} && \mathbb{E}_\mu [f(p_{0:T}(x_{0:T}))] \\ & \text{subject to} && \mathbb{E}_\mu [f_i(p_{0:T}(x_{0:T}))] \leq \epsilon_i \text{ for } i \in [m] \\ & && \mathbb{E}_\mu [p_{0:T}(x_{0:T})] = 1. \end{aligned} \quad (22)$$

The rest follows the same proof steps in the proof of Theorem 1. First, we employ Lyapunov's convexity theorem to show that the epigraph for Problem (22) is non-empty and convex, implying that (22) is strongly dual. Second, we translate the strong duality for Problem (22) to the original problem (21).

To prove strong duality for Problem (11), we introduce some notation as follows. Let $\bar{p}_{0:T, \text{revl}}^*$ be a solution to Problem (21) and denote $\bar{P}_{\text{revl}}^* := D_{\text{KL}}(\bar{p}_{0:T, \text{revl}}^*(\cdot) \parallel q_{0:T}(\cdot; s_q))$. The Lagrangian for Problem (21) is given by $\bar{L}_{\text{revl}}(p_{0:T}(\cdot), \lambda) := \bar{L}_{\text{revl}}(p_{0:T}(\cdot), \lambda)$, and its dual function is given by $\bar{D}_{\text{revl}}(\lambda) := \min_{p_{0:T}} \bar{L}_{\text{revl}}(p_{0:T}(\cdot), \lambda)$. Let an optimal dual variable be $\bar{\lambda}_{\text{revl}}^* \in \arg\max_{\lambda \geq 0} \bar{D}_{\text{revl}}(\lambda)$, and the optimal value of the dual function be $\bar{D}_{\text{revl}}^* := \bar{D}_{\text{revl}}(\bar{\lambda}_{\text{revl}}^*)$. From the strong duality of Problem (21), there exists a pair $(\bar{p}_{0:T, \text{revl}}^*, \bar{\lambda}_{\text{revl}}^*)$ such that

$$\bar{P}_{\text{revl}}^* = \bar{D}_{\text{revl}}^* \text{ or } D_{\text{KL}}(\bar{p}_{0:T, \text{revl}}^*(\cdot) \parallel q_{0:T}(\cdot; s_q)) = \bar{D}_{\text{revl}}(\bar{\lambda}_{\text{revl}}^*).$$

With Assumption 10, this implies that $(\bar{p}_{0:T, \text{revl}}^*, \bar{\lambda}_{\text{revl}}^*)$ is a saddle point of the Lagrangian $\bar{L}_{\text{revl}}(p_{0:T}(\cdot), \lambda)$,

$$\bar{L}_{\text{revl}}(\bar{p}_{0:T, \text{revl}}^*(\cdot), \lambda) \leq \bar{L}_{\text{revl}}(\bar{p}_{0:T, \text{revl}}^*(\cdot), \bar{\lambda}_{\text{revl}}^*) \leq \bar{L}_{\text{revl}}(p_{0:T}(\cdot), \bar{\lambda}_{\text{revl}}^*) \text{ for all } p_{0:T}(\cdot) \text{ and } \lambda \geq 0.$$

By Assumption 6, the score function class \mathcal{S} is expressive enough, any path distribution $p_{0:T}(\cdot)$ can be represented as $p_{0:T}(\cdot; s_p)$ with some $s_p \in \mathcal{S}$; and vice versa. Thus, we can express $\bar{p}_{0:T, \text{revl}}^*(\cdot)$ as $p_0(\cdot; s_{\text{revl}}^*)$ with some $s_{\text{revl}}^* \in \mathcal{S}$. We also note that the dual function $\bar{D}_{\text{revl}}(\lambda)$ in the path and score function spaces are the same. Hence, the optimal dual function for Problem (11) remains to be $\hat{D}_{\text{revl}}(\lambda) = \bar{D}_{\text{revl}}(\lambda)$. Thus, $(s_{\text{revl}}^*, \bar{\lambda}_{\text{revl}}^*)$ is a saddle point of the Lagrangian $\hat{L}_{\text{revl}}(s_p; \lambda) := \bar{L}_{\text{revl}}(p_{0:T}(\cdot; s_p); \lambda)$,

$$\hat{L}_{\text{revl}}(s_{\text{revl}}^*; \lambda) \leq \hat{L}_{\text{revl}}(s_{\text{revl}}^*; \bar{\lambda}_{\text{revl}}^*) \leq \hat{L}_{\text{revl}}(s_p; \bar{\lambda}_{\text{revl}}^*) \text{ for all } s_p \in \mathcal{S} \text{ and } \lambda \geq 0.$$

Therefore, the strong duality holds for Problem (11) in the score function space. \square

C.4. Likelihood Estimation

Lemma 2. Consider $q_0(\cdot)$ and $q_0^i(\cdot)$ to be the distributions of samples for two diffusion models with score functions $s_q(x_t, t)$ and $s_{q^i}(x_t, t)$ respectively. Then the logarithm of the ratio of the probabilities $\log q_0^i(x_0)/q_0(x_0)$ assigned to a given sample x_0 by the two models can be written as:

$$\log q_0^i(x_0) - \log q_0(x_0) \approx \frac{1}{2} \int_0^\infty dt \left(\mathbb{E}_{q(x_t|x_0)} \left[\|\nabla \log q(x_t) - s_q(x_t, t)\|_2^2 \right] - \mathbb{E}_{q(x_t|x_0)} \left[\|\nabla \log q^i(x_t) - s_{q^i}(x_t, t)\|_2^2 \right] \right) \omega_t$$

where ω_t is a time-dependent constant related to the noise schedule of the diffusion process.

Proof. We begin with the setting and notation of (Kong et al., 2023). Let $p(z_\gamma | x)$ be a Gaussian noise channel with

$$z_\gamma = \sqrt{\gamma} x + \epsilon$$

and $\epsilon \sim \mathcal{N}(0, I)$, where γ represents the Signal-to-Noise Ratio (SNR) and $p(x)$ is the unknown data distribution. Then the log probability of the unknown distribution for a given sample x can be written as an integral (Equation 9 from (Kong et al., 2023)):

$$-\log p(x) = \frac{d}{2} \log(2\pi e) - \frac{1}{2} \int_0^\infty d\gamma \left(\frac{d}{1+\gamma} - \text{mmse}(x, \gamma) \right) \quad (23)$$

where we define the point-wise mmse as:

$$\text{mmse}(x, \gamma) := \mathbb{E}_{p(z_\gamma|x)} \left[\|x - \hat{x}^*(z_\gamma, \gamma)\|_2^2 \right] \quad (24)$$

Note that we can write the training loss of a diffusion model trained with samples from $p(x)$ as minimizing the expected point-wise mmse, thus $\hat{x}^*(z_\gamma, \gamma)$ can be viewed as the optimal diffusion model denoiser. With this in mind, we modify the notation slightly to write:

$$\text{mmse}(x_0, t) := \mathbb{E}_{q(x_t|x_0)} \left[\|x_0 - \hat{x}^*(x_t, t)\|_2^2 \right]. \quad (25)$$

Furthermore, from (Luo, 2022) we recall that the denoising diffusion loss and the score matching loss are equivalent up to a time-dependent constant which we denote ω_t allowing us to write:

$$\text{mmse}(x_0, t) = \omega_t \cdot \mathbb{E}_{q(x_t|x_0)} \left[\|\nabla \log p(x_t) - \hat{s}^*(x_t, t)\|_2^2 \right] \quad (26)$$

where $s^*(\cdot, t)$ is the minimizer of the score matching loss $\mathbb{E}_{p(x_0)} \mathbb{E}_{q(x_t|x_0)} \left[\|\nabla \log p(x_t) - \hat{s}(x_t, t)\|_2^2 \right]$. Thus, if we consider two distributions $p_1(x_0), p_2(x_0)$ we can write:

$$\begin{aligned} & \log p_2(x_0) - \log p_1(x_0) \\ &= \frac{1}{2} \int_0^\infty dt (\text{mmse}_1(x_t, t) - \text{mmse}_2(x_t, t)) \\ &= \frac{1}{2} \int_0^\infty dt \left(\mathbb{E}_{q(x_t|x_0)} \left[\|\nabla \log p_1(x_t) - \hat{s}_1^*(x_t, t)\|_2^2 \right] - \mathbb{E}_{q(x_t|x_0)} \left[\|\nabla \log p_2(x_t) - \hat{s}_2^*(x_t, t)\|_2^2 \right] \right) \omega_t. \end{aligned}$$

Again, from (Luo, 2022) we know:

$$\nabla \log p(x_t) = \frac{\sqrt{\bar{\alpha}_t} x_0 - x_t}{1 - \bar{\alpha}_t} \quad (27)$$

where α_t represents the noise schedule, allowing us to compute the integral. \square

D. Algorithms

Here we detail the primal-dual algorithms discussed in Section 3 of the main paper.

D.1. Reverse KL-Constrained Unlearning Algorithm

As discussed in the main paper the algorithm alternates between minimizing the Lagrangian (primal) function, and maximizing the dual function.

Primal step: In the primal step we minimize the Lagrangian:

$$\underset{s_p}{\text{minimize}} \quad \widehat{L}_{\text{rev}}(s_p, \lambda) := D_{\text{KL}}(p_0(\cdot; s_p) \parallel q_0(\cdot; s_q)) - \sum_i \lambda_i (D_{\text{KL}}(p_0(\cdot; s_p) \parallel q_0^i(\cdot; s_q^i)) - b_i) \quad (28)$$

We can estimate the point-wise KL divergences in (28) by sampling trajectories from the model and using the following result (Lemma 2 from (Khalafi et al., 2025)):

Lemma 3 (Point-wise KL divergence). *Assume two score functions $s_p(x, t) = \nabla \log \bar{p}_t(x)$, $s_q(x, t) = \nabla \log \bar{q}_t(x)$, where \bar{p}_t, \bar{q}_t are two marginal densities induced by two forward diffusion processes, with the same noise schedule, starting from initial distributions \bar{p}_0 and \bar{q}_0 , respectively. Then, the point-wise KL divergence between two distributions of the samples generated by running DDIM with s_p and s_q is given by*

$$D_{\text{KL}}(p_0(\cdot; s_p) \parallel q_0(\cdot; s_q)) = \sum_{t=0}^T \tilde{\omega}_t \mathbb{E}_{x_t \sim p_t(\cdot; s_p)} \left[\|s_p(x_t, t) - s_q(x_t, t)\|_2^2 \right] + \epsilon_T \quad (29)$$

where $\tilde{\omega}_t$ is a time-dependent constant, and ϵ_T is a discretization error that depends on the total number of diffusion time steps T and goes to zero as $T \rightarrow \infty$

We can then minimize the weighted sum of the KL divergences using standard gradient-based optimization techniques.

Dual step: In the dual step we update the dual multipliers via $\lambda^+ = \lambda + \eta \partial \widehat{D}_{\text{rev}}(\lambda)$, using the fact that the constraint violations give us a subgradient $\partial \widehat{D}_{\text{rev}}(\lambda)$ of the dual function.

$$\lambda_i^+ = (\lambda_i - \eta_d (D_{\text{KL}}(p_0(\cdot; s_p) \parallel q_0^i(\cdot; s_q^i)) - b_i))_+ \quad (30)$$

A simple observation regarding the dual update is that if a constraint is not satisfied, we increase the corresponding multiplier.

D.2. Forward KL-Constrained Unlearning Algorithm

Similar to the reverse KL algorithm, we alternate between minimizing the Lagrangian (primal) function, and maximizing the dual function.

Primal step: In the primal step we minimize the Lagrangian:

$$\underset{s_p}{\text{minimize}} \quad \widehat{L}_{\text{fw}}(s_p, \lambda) := D_{\text{KL}}(q_{0:T}(\cdot; s_q) \parallel p_{0:T}(\cdot; s_p)) - \sum_i \lambda_i (D_{\text{KL}}(q_{0:T}^i(\cdot; s_q^i) \parallel p_{0:T}(\cdot; s_p)) - b_i). \quad (31)$$

The forward KL divergence $D_{\text{KL}}(q_{0:T}(\cdot; s_q) \parallel p_{0:T}(\cdot; s_p))$ is the standard diffusion model training objective which we can estimate using samples from the distribution q :

$$D_{\text{KL}}(q_{0:T}(\cdot; s_q) \parallel p_{0:T}(\cdot; s_p)) = \mathbb{E}_{x_0 \sim q} \mathbb{E}_{q(x_t|x_0)} \left[\|s_p(x_t, t) - \nabla \log q(x_t, t)\|_2^2 \right] + \text{constant}. \quad (32)$$

The first term on the RHS of (32) is the standard score matching objective with samples from the distribution q . The Lagrangian then becomes:

$$\widehat{L}_{\text{fw}}(s, \lambda) = \mathbb{E}_{x_0 \sim q} \mathbb{E}_{q(x_t|x_0)} \left[\|s_p(x_t, t) - \nabla \log q(x_t, t)\|_2^2 \right] - \sum_i \lambda_i (\mathbb{E}_{x_0 \sim q_i} \mathbb{E}_{q(x_t|x_0)} \left[\|s_p(x_t, t) - \nabla \log q^i(x_t, t)\|_2^2 \right] - b_i) \quad (33)$$

which we can minimize using standard gradient-based optimization techniques using samples from q and q^i .

Dual step: In the dual step, similar to the reverse KL algorithm we simply update the dual multipliers dependent on the constraint violations:

$$\lambda_i^+ = (\lambda_i - \eta_d (D_{\text{KL}}(q_{0:T}^i \| p_{0:T}(\cdot; s_p)) - b_i))_+. \quad (34)$$

D.3. Likelihood-Constrained Unlearning Algorithm

The algorithm similarly alternates between primal and dual steps.

Primal step: For the primal step we need to minimize the Lagrangian:

$$\underset{s_p}{\text{minimize}} \widehat{L}_{\text{revl}}(s_p, \lambda) := D_{\text{KL}}(p_{0:T}(\cdot; s_p) \| q_{0:T}(\cdot; s_q)) + \sum_i \lambda_i (\mathbb{E}_p[q^i(x_0)] - \epsilon_i). \quad (35)$$

We can treat the likelihood of the unlearning distribution $q^i(x_0)$ as a reward function that can be evaluated for any given sample using a diffusion model that samples from the distribution $q^i(\cdot)$. We discuss this in detail in Lemma 2 the result of which we summarize here:

$$\log q_0^i(x_0) - \log q_0(x_0) \approx \frac{1}{2} \sum_{t=0}^T \left(\mathbb{E}_{q(x_t|x_0)} \left[\|\nabla \log q(x_t) - s_q(x_t, t)\|_2^2 \right] - \mathbb{E}_{q(x_t|x_0)} \left[\|\nabla \log q^i(x_t) - s_{q^i}(x_t, t)\|_2^2 \right] \right)$$

Note we have further approximated the integral in Lemma 2 with a sum over discrete diffusion time steps.

Now that we have a way to estimate the likelihood, the primal minimization becomes the well-known problem of aligning a diffusion model with a penalty function with the reward being the unlearning likelihood $q(x_0)$ in this case. We can use the following result from (Fan et al., 2023b) to compute the gradient of the expected likelihood:

$$\nabla_{\theta} \mathbb{E}_{p_{\theta}(x_0)}[q(x_0)] = \mathbb{E}_{p_{\theta}(x_{0:T})} \left[q(x_0) \sum_{t=1}^T \nabla_{\theta} \log p_{\theta}(x_{t-1} | x_t) \right] \quad (36)$$

which allows us to minimize the Lagrangian in (35) using gradient-based optimization.

Dual step: For the dual steps we update the dual variables via:

$$\lambda_i = (\lambda_i + \eta_d (\mathbb{E}_p[q^i(x_0)] - \epsilon_i))_+. \quad (37)$$

E. Additional Implementation Details

Here we provide some additional implementation details regarding the experiments discussed in Section 4. All experiments were run on an NVIDIA B200 GPU.

E.1. Likelihood-Constrained Unlearning

E.1.1. GAUSSIAN MIXTURE UNLEARNING

For the Gaussian mixture unlearning experiments we trained diffusion models with ResNet (He et al., 2015) networks as denoisers. We used a DDPM (Ho et al., 2020) scheduler with $T = 100$ diffusion time steps. For the reverse KL baseline, we set the dual update learning rate to 0 to essentially minimize the Lagrangian with a fixed multiplier λ , and we ran for different values of λ ranging from 0.01 to 1.0. For the Likelihood constrained approach we ran for different constraint thresholds ranging from 0.05 to 1.0. The likelihood is normalized such that the expected likelihood of the unlearn concept is approximately 1.0 when evaluated over samples of the initial pretrained model.

The constraint thresholds are chosen by multiplying the expected likelihood of the unlearning evaluated on samples from the pretrained model, by a factor.

We also note that since Lemma 2 allows the estimation of ratio of likelihoods, we consider the likelihood of the distribution that is proportional to $q^i(x)/q_{pre}(x)$, where $q_{pre}(x)$ is the pretrained model.

The full list of important hyperparameters is listed in Table 1.

Table 1. Hyperparameters for constrained likelihood vs fixed weight reverse KL diffusion unlearning experiments. Used for producing the results in Figure 4.

Category	Hyperparameter	Value
Training	Number of epochs (N_{epochs})	100
	Batches per epoch	1
	Primal learning rate (η_p)	2×10^{-4}
	Dual learning rate (η_d)	0/0.1
	Primal steps per dual step	5
	Primal batch size	128
	Gradient accumulation steps	1
	Optimizer	AdamW
	LR schedule	Cosine w/ warmup
Diffusion	Number of diffusion steps (T)	100
	Inference steps	10
	Noise scheduler	DDPM
	noise scheduler β_{start}	1×10^{-4}
	noise scheduler β_{end}	2×10^{-2}
Model architecture	Model	ResNet Diffusion Model
	Input dimension (x_{dim})	2
	Hidden dimension (h_{dim})	128
	Embedding dimension	32
	Number of layers	4
	Widen factor	2
	Time embedding	Learned

E.1.2. CONCEPT UNLEARNING WITH STABLE DIFFUSION

We use a distilled version of Stable Diffusion v1.4 for these experiments. We fine-tune the parameters of LoRA adapters (Hu et al., 2021) added to the cross-attention layers in the diffusion U-Net following the esd-x approach from (Gandikota et al., 2023).

For computing KID to the retain distribution, we construct a retain set by sampling images from the pretrained model and keeping those whose likelihood/CLIP score for the unlearn concept is lower than a certain threshold i.e., samples in which

Table 2. Hyperparameters for fixed weight concept erasure baseline (unconstrained) and Likelihood-constrained unlearning sweeps. Used for producing the results in Figure 5.

Category	Hyperparameter	Value
General	Base model	huggingface/nota-ai/bk-sdm-tiny
	Mixed precision	FP16
	Number of epochs	30
	Training batch size	8
	Gradient accumulation steps	1
	Learning rate	5×10^{-4}
	Optimizer	Adam
	LoRA rank (r)	8
	LoRA scaling (α)	8
	Primal steps per dual step	2
Diffusion / sampling	Inference steps (train)	20
	Noise samples per timestep for likelihood estimation	8
	Expected Likelihood Evaluation batch size	16
Retain set generation	Number of retain images	1000
	clip threshold factor	0.95
	likelihood threshold factor	1.0
Unconstrained Baseline	Dual learning rate (η_d)	0
	Fixed multiplier sweep values	{0.01, 0.1, 0.2, 0.3, 0.5, 0.75, 1.0, 1.5, 2.0, 2.5, 3.0, 5.0}
Likelihood-constrained	Dual learning rate (η_d)	10^{-3}
	Constraint threshold (ϵ)	$\bar{\ell} \times \beta$
	threshold scaling factors (β)	{0.70, 0.71, ..., 0.95}

the unlearn concept does not appear. The threshold is the expected likelihood/CLIP over samples from the pretrained model multiplied by a factor.

Similarly the likelihood constraint thresholds were chosen by multiplying the expected likelihood of the unlearn concept for the pretrained model ($\bar{\ell}$) by factors smaller than 1. We highlight that since Lemma 2 allows the estimation of ratio of likelihoods, we consider the likelihood of the distribution that is proportional to $q_{pre}(x, c_i)/q_{pre}(x)$, where $q_{pre}(x)$ is the unconditional pretrained distribution of Stable Diffusion and $q_{pre}(x, c_i)$ is conditioned on the unlearning concept c_i .

We also note that, unlike the Gaussian case where the likelihood constraint is reliably satisfied regardless of threshold level, for Stable Diffusion, optimizing the expected likelihood is challenging as the loss sometimes diverges. In Figure 5 we report only the convergent runs.

The full list of important hyperparameters is listed in Table 2.

E.2. Forward KL-Constrained Unlearning

For forward KL constrained unlearning we unlearn specific samples from a model pretrained on the CelebA-HQ dataset. Our implementation is based on modifying that of (Alberti et al., 2025), namely, SISS without importance sampling, to incorporate dual updates.

For the unconstrained baseline, we use the same fixed multiplier λ for all the unlearning samples. We sweep λ over a range from 0.05 to 0.5 to observe SSCD values from 1.0 (almost no unlearning) to around 0.3 which by visual inspection represents complete unlearning of the memorized sample.

We compute the SSCD score between two images: the original image sample we wish to unlearn from the model, and the image obtained by adding noise to the unlearning sample up to time step $t = 200$ and then denoising using the model to see how closely it can recover the original sample. This comparison is visually represented in Figure 2, Left.

The full list of important hyperparameters is listed in Table 3.

Table 3. Hyperparameters for forward KL constrained data unlearning experiments on CelebA-HQ data samples. Used to obtain the results in Figures 2, 3.

Category	Hyperparameter	Value
General	Dataset	CelebA-HQ (256 × 256)
	Pretrained model	google/ddpm-celebahq-256
	Diffusion time steps T	1000
Training	Epochs	200
	Train batch size	2
	Gradient accumulation steps	16
	Effective batch size	32
	Learning rate (η_p)	5×10^{-6}
	Dual learning rate (η_d)	5×10^{-2}
	Optimizer	Adam
Constraints	Constraint type	Forward KL divergence
	Threshold values	{0.0025, 0.005, 0.0075, 0.010, 0.015, 0.02, 0.03}
Unconstrained	fixed multiplier sweep values	{0.01, 0.015, 0.02, 0.025, 0.030, 0.035, 0.04, 0.045, 0.05}

E.3. Reverse KL-Constrained Unlearning

Dual-Only Algorithm. For the reverse KL experiments we modify the algorithm slightly. We can skip the primal step in the algorithm discussed in Appendix D.1, by observing the fact that we already know from section 2.1 that for a given λ the optimal primal minimizer is given by

$$q_u^\dagger(\cdot; \lambda) = \frac{1}{Z_u^\dagger(\lambda)} \frac{(q(\cdot))^{1/(1-\mathbf{1}^\top \lambda)}}{\prod_{i=1}^m (q_u^i(\cdot))^{\lambda_i}}.$$

Since we have access to the score functions of $q(\cdot)$, and the scores of the individual unlearning distributions $q^i(\cdot)$, we can bypass the primal step and directly do the dual updates in (30).

This dual-only algorithm allows us to find the optimal multipliers λ^* that satisfy the unlearning reverse KL constraints. If needed, the model weights can then be fine-tuned so that the distribution of generated samples fits $q_u^\dagger(\cdot; \lambda^*)$. Khalafi et al. (2025) propose a similar dual-only algorithm to find the optimal weights for diffusion model composition.

To choose the constraint thresholds, we first estimate the reverse KL between the pretrained model, and each of the individual unlearning concepts. Then we set the threshold as the average of these KL divergences, multiplied by a factor.

For the fixed equal weight, unconstrained baseline, we choose all dual multipliers to be equal, and across runs we sweep the multiplier values so that the sum of the dual multipliers, $\mathbf{1}^\top \lambda$, goes from 0 to 0.95. Recall from Section 2.1 that to avoid degenerate solutions we need $\mathbf{1}^\top \lambda < 1$.

Table 4. Hyperparameters for reverse KL constrained unlearning experiments with dual-only algorithm. Used to obtain the results in Figure 6.

Category	Hyperparameter	Value
General	Base model	Stable Diffusion v1.4
	Prompt structure	1 retain prompt + multiple unlearn prompts
	Example prompt	“a photo of a dog, a golden retriever, a german shepherd, a brown dog”
Training	Epochs (heuristic b sweep)	75
	Training batch size	8
	Primal learning rate	10^{-4}
	Dual learning rate (η_d)	0.02
Diffusion / sampling	Guidance scale	7.5
	Inference steps	10
Dual / constraint settings	KL type	Pathwise
	Multiplicative threshold factor	{2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15}