Provably Safe Generative Sampling with Constricting Barrier Functions

Anonymous authors
Paper under double-blind review

Abstract

Flow-based generative models, such as diffusion models and flow matching models, have achieved remarkable success in learning complex data distributions. However, a critical gap remains for their deployment in safety-critical domains: the lack of formal guarantees that generated samples will satisfy hard constraints. We propose a safety filtering framework that acts as an online shield for any pre-trained generative model. Our key insight is to cooperate with the generative process rather than override it. We define a constricting safety tube that is relaxed at the initial noise distribution and progressively tightens to the target safe set at the final data distribution, mirroring the coarse-to-fine structure of the generative process itself. By characterizing this tube via Control Barrier Functions (CBFs), we synthesize a feedback control input through a convex Quadratic Program (QP) at each sampling step. As the tube is loosest when noise is high and intervention is cheapest in terms of control energy, most constraint enforcement occurs when it least disrupts the model's learned structure. We prove that this mechanism guarantees safe sampling while minimizing the distributional shift from the original model at each sampling step, as quantified by the KL divergence. Our framework applies to any pre-trained flow-based sampling scheme requiring no retraining or architectural modifications. We validate the approach across constrained image generation, physically-consistent trajectory sampling, and safe robotic manipulation policies, achieving 100% constraint satisfaction while preserving semantic fidelity.

1 Introduction

Flow-based generative models such as diffusion models (Ho et al., 2020; Song et al., 2021), flow matching (Lipman et al., 2023), and continuous normalizing flows (Papamakarios et al., 2021) have redefined the state-of-the-art in learning complex, high-dimensional distributions. By transforming a simple prior noise distribution into a structured data distribution through a sequence of infinitesimal steps, these models excel in tasks such as molecular design (Weiss et al., 2023), high-fidelity image synthesis (Ho et al., 2022), and control policies for robots (Chi et al., 2025). The deployment of these models increasingly requires constraint satisfaction across diverse applications. In content generation, constraints ensure human alignment, which requires filtering harmful content in images (Schramowski et al., 2023) or preventing toxic text generation (Li et al., 2025). In safety-critical physical systems such as robotics (Janner et al., 2022; Chi et al., 2025) or autonomous navigation (Liao et al., 2025), constraints encode inviolable physical laws or safety specifications such as dynamical feasibility, joint limits, collision avoidance, smoothness, etc.

Traditional soft guidance techniques, such as classifier-based, classifier-free (Dhariwal & Nichol, 2021; Ho & Salimans, 2021) or reward-weighted guidance (Yuan et al., 2023), merely act as probabilistic incentives. While they bias the model toward desired regions, they cannot provide provable guarantees of feasibility. Conversely, projection-based methods can guarantee safety by formally projecting completed samples onto a safe manifold (Fishman et al., 2023; Utkarsh et al., 2025), yet they often suffer from significant computational overhead and introduce large distributional shifts.

To address this gap, we propose a guidance scheme motivated by safety filtering in control theory. By leveraging Control Barrier Functions (CBFs), we synthesize a feedback control input through a Quadratic

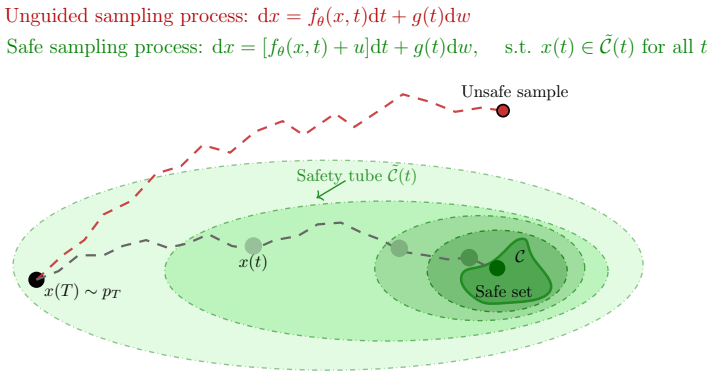


Figure 1: *Constrained generative sampling with a constricting safety tube.* The flow-based sampling process transforms noise sample $x(T)$ into data $x(0)$. *Unconstrained sampling process* can produce unsafe samples that violate constraints, landing outside the required safe set. Our control barrier function-guided sampling augments the generative dynamics with feedback control inputs u that maintain the sample within the constricting safety tube $\tilde{\mathcal{C}}(t)$ (green region) throughout the sampling process, guaranteeing safe samples and minimal perturbation of the learned sampling process.

Program (QP) that aims to retain the fidelity of the original model. Our key insight is the use of a constricting safety tube $\tilde{\mathcal{C}}(t)$ that is relaxed at the high-noise regime ($t = T$) and progressively constricts to the target safe set $\mathcal{C}$ at the data distribution ($t = 0$). As illustrated in Figure 1, this ensures that the final sample belongs to the safe set while formally bounding the distributional shift between the original model and the safe distribution. We frame the guidance problem as a problem of control synthesis. The goal is to inject a feedback control input u into the sampling process to render the safety tube $\tilde{\mathcal{C}}(t)$ invariant, ensuring that trajectories provably remain within the safety tube for the entire duration. CBFs provide the mathematical framework needed to synthesize such control inputs with formal safety certificates regarding the membership of $x(t)$ in $\tilde{\mathcal{C}}(t)$. Our framework ensures that the sample at the final sampling step $x(0)$ lies within the safe set $\mathcal{C}$ which satisfies the required constraints. At each sampling step, we solve a constrained optimal control problem that minimizes control effort to preserve distributional fidelity while enforcing a CBF constraint that guarantees $x(t) \in \tilde{\mathcal{C}}(t)$ for all $t \in [0, T]$.

Unlike prior CBF-based approaches (Xiao et al., 2023; Yang et al., 2025; Dai et al., 2025) that employ prescribed-time convergence or post-hoc trajectory correction, our method enforces strict invariance of a constricting safety tube that tightens during the sampling process, providing safety guarantees on sampling. Our main contributions are:

1. *Provably safe sampling:* For any closed and bounded set $\mathcal{C}$, we prove that our guidance mechanism based on CBFs ensures that the final sample $x(0) \in \mathcal{C}$. Importantly, we make no assumptions about the convexity of the safe set.

2. *Cooperation with the generative process:* Our constricting safety tube mirrors the coarse-to-fine structure of flow-based sampling, concentrating constraint enforcement in the high-noise regime where interventions are distributionally cheap. This ensures that the model retains full authority over the semantic structure and fine details. We prove that our minimum-norm control minimizes the per-step contribution to the KL divergence between the safe and original distributions.

3. *Modular guidance*: Our guidance scheme can be applied to any pre-trained flow-based generative model at sampling time, requiring no retraining or architectural modifications.

We validate our approach on three different experiments: simulation of nonlinear physics, constrained image generation, safety-critical planning in robotics.

2 Related Work

The literature on enforcing constraints in generative models is rapidly growing, typically bifurcated into methods that provide probabilistic incentives and those that formally enforce hard constraints.

Soft guidance and probabilistic steering. Soft guidance methods generally modify the sampling process by adding a penalty or a steering term that biases the model toward desired regions of the data distribution. Classifier guidance (Dhariwal & Nichol, 2021) utilize the probabilistic confidence from a classifier, or implicitly compute confidence in classifier-free guidance Ho & Salimans (2021) to steer the score function towards specific attributes or labels. Similarly, reward-guided methods Yuan et al. (2023) prioritize trajectories that maximize rewards, while gradient-based approaches (Guo et al., 2024) steer sampling via optimization objectives. While these methods effectively increase the likelihood of constraint satisfaction, they do not provide formal guarantees of admissibility. They lack a mechanism to explicitly ensure safe sampling. They are fundamentally probabilistic and may still produce samples that violate safety guidelines, making them unsuitable for safety-critical hardware where failure is impermissible.

Strict guidance with projection. To address the need for deterministic certificates, several hard enforcement frameworks have recently emerged, largely focused on projection. Projection-based approaches (Fishman et al., 2023; Utkarsh et al., 2025; Römer et al., 2024; Gadginmath & Pasqualetti, 2025; Zampini et al., 2025) project unconstrained samples onto a valid manifold of samples at each sampling step. Classically, constraining a stochastic process to a domain is framed as a Skorokhod problem, which utilizes reflected SDEs (Lou & Ermon, 2023) to push the process back into the set whenever it reaches the boundary of the domain. While mathematically elegant, these approaches can be computationally intensive for complex, non-convex sets and may introduce distributional shifts when the reflection direction is not well-aligned with the learned drift.

CBF-based guidance in sampled robotic planning and control. A recent line of literature in sampling based robotic planning and control have increasingly turned to CBFs (Mizuta & Leung, 2024; Dai et al., 2025; Xiao et al., 2023; Yang et al., 2025) to transform soft heuristics into formal safety certificates. Mizuta & Leung (2024) use a CBF-CLF framework to define a reward function, which is in turn used for reward-based guidance. Both Dai et al. (2025) and Xiao et al. (2023) employ prescribed-time CBF mechanisms with singular class-$\mathcal{K}$ functions that enforce convergence to the safe set by a deadline, requiring high-gain feedback that can lead to aggressive steering. Yang et al. (2025) adopts a prediction-correction architecture that generates a candidate path in latent space and then corrects it in a separate, post-hoc phase. In contrast, our method enforces strict invariance throughout the entire sampling process by utilizing a constricting safety tube. This ensures the particle remains safe at every integration step, rather than only at the final time. By using a minimum-norm control synthesis objective, we maintain maximal model fidelity without the need for high-gain steering. Our approach integrates the safety barrier directly into the sampling dynamics as a control intervention. We prove that our method minimizes the per-step contribution to the KL divergence between the safe and learned distributions.

3 Preliminaries: Flow-based sampling and CBFs

Here we introduce essential mathematical preliminaries. We introduce generative sampling from the perspective of dynamical systems and the use of CBFs for safe control.

3.1 Flow-based sampling as a dynamical system

We formalize generative sampling as the trajectory of an autonomous system that transforms a prior noise distribution into a target data distribution. We define $t \in [0, T]$ as the sampling time, where $p(T)$ is the initial noise distribution at $t = T$, and $p(0)$ is the final data distribution at $t = 0$. The sampling process is governed by a Stochastic Differential Equation (SDE):

$$\mathrm{d}x = f_\theta(x, t)\mathrm{d}t + g(t)\mathrm{d}w, \tag{1}$$

where the sample is $x(t) \in \mathbb{R}^n$, and $f_\theta(x,t)$ is the drift vector field. The noise schedule $g(t)$ is time-dependent, and $\mathrm{d}w$ is a reverse-time Wiener process (Song et al., 2021), reflecting that t decreases from T to 0 during sampling. We are able to capture a large family of flow-based generative models using this notation. For example, in score-based diffusion models (Song et al., 2021), $f_\theta(x,t) = -\frac{1}{2}x - g^2(t)\nabla_x \log p(x,t)$ where the score $\nabla_x \log p(x,t)$ is approximated by a neural network. In flow matching (Lipman et al., 2023), $f_\theta = \mathrm{NN}_\theta(x,t)$ directly learns the velocity field that generates optimal transport paths between noise and data distributions. For deterministic frameworks like flow matching, the noise term vanishes, i.e. $g(t) = 0$, reducing the dynamics to an Ordinary Differential Equation (ODE).

While the SDE (1) describes individual sample paths, it induces a corresponding evolution of the population's probability density $p(x,t)$ as t decreases from T to 0. This progression has an important structural property: in the high-noise regime ($t \approx T$), the broad support of $p(x,t)$ means the model establishes only global, coarse structure. As $t \to 0$ and the distribution concentrates toward the data distribution $p(0)$, the model progressively resolves finer details. We refer to this as the *coarse-to-fine* structure of flow-based sampling, and our constricting safety tube in section 4 is designed to mirror this progression. This coarse-to-fine progression has a direct consequence for constrained sampling: interventions applied during the high-noise regime, where the model has not yet committed to fine-grained structure, are less disruptive than those applied near $t = 0$ when the sample has nearly converged.

3.2 Control Barrier Functions for Set Invariance

Consider a system whose state $x(t) \in \mathbb{R}^n$ evolves as t decreases from T to 0, governed by:

$$\mathrm{d}x = f(x,u)\,\mathrm{d}t, \tag{2}$$

where $u \in \mathbb{R}^m$ is the control input. Since t decreases during sampling, $\mathrm{d}t < 0$ throughout, and $f(x,u)$ represents the drift of the reverse-time dynamics, consistent with the sampling process (1). Assume that we are given a set $\mathcal{C} \subset \mathbb{R}^n$ of "safe" states, given by the superlevel set of a continuously differentiable function $h : \mathbb{R}^n \to \mathbb{R}$:

$$\mathcal{C} = \{x \in \mathbb{R}^n \mid h(x) \geq 0\}. \tag{3}$$

We seek to design a feedback controller that ensures the state remains in $\mathcal{C}$ throughout the sampling process. This is accomplished using a feedback controller design based on control barrier functions Ames et al. (2016; 2019). Formally, we say $\mathcal{C}$ is *reverse invariant*[1] if, for any terminal condition $x(T) \in \mathcal{C}$, the resulting trajectory satisfies $x(t) \in \mathcal{C}$ for all $t \in [0,T]$ as t decreases from T to 0. Equivalently, if $h(x(T)) \geq 0$, we need that $h(x(t)) \geq 0$ for all $t \in [0,T]$.

We say that h is a *reverse-time control barrier function*[2] if for all $x \in \mathcal{C}$, there exists a $u \in \mathbb{R}^m$

$$\nabla h(x) \cdot f(x,u) \leq \gamma(h(x)). \tag{4}$$

where γ is a class-$\mathcal{K}$ function, i.e. γ is continuous, strictly increasing, and $\gamma(0) = 0$. If h is a reverse-time CBF, then any feedback controller $u(x,t)$ that satisfies (4) renders $\mathcal{C}$ reverse invariant.

The condition, (4) may be understood intuitively as follows: to maintain reverse invariance of $\mathcal{C}$ we require that for any trajectory $x(t)$ that reaches the boundary $\partial\mathcal{C}$, i.e. $h(x(t)) = 0$, there exists an input $u \in \mathbb{R}^m$ such that $\frac{dh}{dt} \leq 0$. Since t is decreasing, $\frac{dh}{dt} \leq 0$ implies that h is increasing. This ensures that, in reverse time, h increases, and the trajectory remains on the boundary or returns to the interior of $\mathcal{C}$. The CBF condition (4) enforces this condition gradually throughout the entire safe set: if $x(t)$ is instead in the interior of $\mathcal{C}$, $\gamma(h(x(t))) > 0$, permitting inputs such that $\frac{dh}{dt} > 0$, while continuity of γ still guarantees that $\frac{dh}{dt} \to 0$ and $x \to \partial\mathcal{C}$.

[1]Note that standard literature in safety-critical control deals with control systems that evolve forward in time, hence the results derive conditions for forward invariance. Since we are controlling generative sampling dynamics that evolve backward in time, we instead consider the problem of enforcing reverse invariance. Our presentation closely follows Ames et al. (2019), however the CBF condition (4) is modified to be consistent with the change in direction of time.

[2]For brevity, we henceforth refer to reverse-time control barrier functions simply as CBFs since we only deal with systems evolving backward in time.

4 Guidance of flow-based models via constricting CBFs

Having established the sampling dynamics (1) and the CBF framework for safe control, we address the central question: how can we enforce hard constraints on the output of a pre-trained generative model without modifying the model itself? The key observation is that by introducing a control input to the sampling SDE (1), we arrive at a control-affine structure for the control systems in Section 3.2. This additive control input u can be designed to steer sampling trajectories toward the safe set while preserving the learned distribution. This is a control synthesis problem with the following guided sampling process:

$$\mathrm{d}x = [f_\theta(x,t) + u(x,\xi,t)]\mathrm{d}t + g(t)\mathrm{d}w, \quad t \in [0,T], \tag{5}$$

where f_θ is the unconstrained drift learned by the generative model, $g(t)$ is the noise schedule, $\mathrm{d}w$ is a Wiener process, and $u(x,\xi,t)$ is the feedback control that depends on the current state x and the realized noise[3] $\xi = \mathrm{d}w/\mathrm{d}t$. The challenge is to design u so that it provides formal safety guarantees while remaining minimally invasive, intervening only when necessary and as little as possible.

Note that our proposed feedback controller depends both on the state of the system and the noise process. Although this assumption may be restrictive in settings where the feedback controller regulates a physical system, and the noise cannot be directly measured, in our application we have full knowledge of the noise process since the system is being simulated on a computer. This assumption allows us to ensure that reverse invariance holds exactly for the stochastic system, whereas other approaches that consider feedback controllers depending only the state do not have such guarantees (c.f. Remark 1).

4.1 CBFs for flow-based generative sampling

A fundamental challenge in generative sampling is that the initial prior noise $x(T)$ is drawn from a distribution such as the standard Gaussian. This means that $x(T)$ almost certainly resides outside the target safe set $\mathcal{C}$. Traditional CBF formulations for static sets would require an immediate, high-energy intervention to push $x(t)$ close to $\mathcal{C}$, which would violate the model's generative intent and produce samples that no longer represent the learned distribution. Our framework is motivated by the need for minimal interventions and a smooth approach towards $\mathcal{C}$. We achieve this by defining a constricting barrier $\tilde{h}(x,t) = h(x) + \epsilon(x(T),t)$. This is a time-varying barrier that constricts with the sampling process. We characterize safety through a constricting superlevel tube $\tilde{\mathcal{C}}(t)$ with respect to the constricting barrier $\tilde{h}$ as,

$$\tilde{\mathcal{C}}(t) = \{x \in \mathbb{R}^n \mid \tilde{h}(x,t) \geq 0\}. \tag{6}$$

We design $\tilde{\mathcal{C}}(t)$ such that initially, the set $\tilde{\mathcal{C}}(T)$ is sufficiently relaxed to contain the noise sample $x(T)$, and at the end of sampling, it recovers the target set $\mathcal{C}$, i.e., $\tilde{\mathcal{C}}(0) = \mathcal{C}$. Since the sampling process evolves in reverse time from $t = T$ to 0, the safety tube $\tilde{\mathcal{C}}(t)$ must remain occupied by $x(t)$ as t decreases, a property we term *reverse invariance*. As derived in Section 3.2, the reverse-time CBF condition requires bounding $\nabla\tilde{h} \cdot f$ from above. We design the constricting barrier function through a time-varying control barrier function.

Definition 1 (Constricting barrier function) *Given a CBF $h : \mathbb{R}^n \to \mathbb{R}$ and a terminal condition $x(T)$, a constricting barrier function is $\tilde{h}(x,t) = h(x) + \epsilon(x(T),t)$, where ϵ is a C^1 function satisfying:*

1. ***Initial feasibility:*** $\epsilon(x(T),T) \geq \max(0, -h(x(T)))$.
2. ***Recovery of target set:*** $\epsilon(x(T),0) = 0$.
3. ***Monotone constriction:*** $\partial\epsilon/\partial t \geq 0$ *for all* $t \in [0,T]$. *Since t decreases from T to 0 during sampling, this means ϵ decreases as sampling progresses, constricting $\tilde{\mathcal{C}}(t)$ toward $\mathcal{C}$.*

The associated constricting safety tube is $\tilde{\mathcal{C}}(t) = \{x \in \mathbb{R}^n \mid \tilde{h}(x,t) \geq 0\}$.

[3] We say that $\xi : [0,T] \to \mathbb{R}^n$ is the *realized noise* to a Wiener process w if $\int_0^t \xi(\tau)d\tau = w(t)$ for all $t \in [0,T]$. Although w is not continuously differentiable, ξ may be thought of as a formal derivative the Wiener process.

The following result shows how to ensure that the sampling process (5) can be made invariant to the safety tube $\tilde{\mathcal{C}}(t)$ using a constricting barrier function.

Theorem 4.1 (Reverse invariance) *Let $\tilde{h}$ be a constricting barrier function and γ a class-$\mathcal{K}$ function. For each noise realization $\{\xi(t)\}_{t\in[0,T]}$ as t decreases from T to 0, if there exists a control $u(x,\xi,t)$ such that*

$$\nabla_x \tilde{h}(x,t)\cdot(f_\theta(x,t)+u(x,\xi,t)+g(t)\xi(t))+\frac{\partial \tilde{h}(x,t)}{\partial t}\leq \gamma(\tilde{h}(x,t)), \tag{7}$$

then the guided sampling process ensures $x(t)\in\tilde{\mathcal{C}}(t)$ for all $t\in[0,T]$, and in particular $x(0)\in\mathcal{C}$.

We refer the reader to Appendix A.1 for a formal proof. Theorem 4.1 establishes that any control u satisfying the CBF condition (7) guarantees $x(0)\in\mathcal{C}$, regardless of the convexity of $\mathcal{C}$, the architecture of f_θ, or the location of the initial noise sample $x(T)$. The key mechanism is the interplay between the constriction and the class-$\mathcal{K}$ function. At the onset of sampling ($t\approx T$), $\tilde{h}=h+\epsilon$ is large due to the relaxation ϵ. As sampling progresses, $\epsilon\to 0$ and $\tilde{h}\to h$, so $\gamma(\tilde{h})$ shrinks and the constraint tightens toward the standard CBF condition for $\mathcal{C}$ guided by the feedback control $u(x,\xi,t)$. By this point, the learned model f_θ has already resolved the trajectory close to $\mathcal{C}$, and the control u needs only to enforce the final constraint boundary. Many fixed-time convergence schemes rely on singular class-$\mathcal{K}$ functions that blow up as $t\to 0$ (Xiao et al., 2023; Dai et al., 2025). In contrast, our C^1 constriction mimics the coarse-to-fine generation scheme of typical flow-based models, ensuring that any feasible control remains bounded and minimally invasive.

Remark 1 (Comparison to other approaches to extending CBFs to stochastic systems) *Unlike works that consider designing feedback controllers that enforce safety of stochastic dynamical systems (Clark, 2021), our approach assumes that the feedback controller has knowledge of the state as well as the noise. While this assumption is more restrictive, it allows us to obtain stronger guarantees. In particular, with our design approach, we can enforce invariance of the constricting safety tube exactly, whereas by So et al. (2023), invariance does not exactly or almost surely for the CBFs considered in Clark (2021).*

Theorem 4.1 provides safety given the noise realization $\xi(t)$. At each step, we first observe the noise realization $\xi(t)\sim\mathcal{N}(0,I)$, then synthesize $u=u(x,\xi,t)$ to ensure safety given the observed noise. This paradigm is conditioned on the observed noise at each step, providing deterministic safety for each trajectory realization rather than safety in expectation. While Gaussian noise has unbounded support and extreme realizations could theoretically require large control effort, our experiments (Section 5) demonstrate reliable feasibility across diverse applications. Alternative formulations that use model predictive control require stochastic CBFs (Prajna et al., 2004) to predict the behavior of model and the noise. Here, chance constraints can provide formal high-probability bounds by incorporating safety margins proportional to the noising scheme $g(t)$. We defer such extensions to future work.

We now discuss candidates for the constriction scheme $\epsilon(x(T),t)$. Let $\epsilon_0=\max(0,-h(x(T)))$ denote the initial constraint violation. A key practical advantage of Definition 1 is that its conditions are mild enough to be satisfied by a broad family of simple, closed-form functions. We present three such candidates: (1) *Linear*: $\epsilon(x(T),t)=\epsilon_0\cdot(t/T)$ with constant constriction rate; (2) *Exponential*: $\epsilon(x(T),t)=\epsilon_0\cdot(e^{\lambda t/T}-1)/(e^{\lambda}-1)$ for $\lambda>0$, which front-loads relaxation when $\lambda>1$, providing aggressive early intervention and gentler later refinement; (3) *Polynomial*: $\epsilon(x(T),t)=\epsilon_0\cdot(t/T)^p$ for $p\geq 1$, which back-loads constriction when $p>1$, offering slower initial relaxation and faster final convergence. All schemes satisfy Definition 1 by construction, and the shape parameter (λ or p) gives practitioners direct control over when along the sampling trajectory the constriction pressure is concentrated.

Minimum-norm control synthesis. To ensure low-energy interventions that preserve model fidelity, we synthesize the control as:

$$\begin{aligned}\min_u \quad & \frac{1}{2}\|u\|^2\\ \text{s.t.} \quad & \nabla_x \tilde{h}(x,t)\cdot(f_\theta(x,t)+u+g(t)\xi(t))+\frac{\partial \epsilon(x(T),t)}{\partial t}\leq \gamma(\tilde{h}(x,t)),\end{aligned} \tag{8}$$

where the optimization is performed for the current state-noise pair (x,ξ) at time t, yielding $u^*(x,\xi,t)$.

Remark 2 (Feasibility of the QP (8)**)** *The above QP is a minimum-norm problem with a single linear constraint of the form $a^\top u \leq b$, where $a = \nabla_x \tilde{h}(x,t)$ and $b = \gamma(\tilde{h}) - \nabla_x \tilde{h} \cdot (f_\theta + g\xi) - \frac{\partial \epsilon}{\partial t}$. When $b \geq 0$, the uncontrolled drift already satisfies the CBF condition and $u^* = 0$. When $b < 0$, the QP admits the closed-form solution $u^* = (b/\|a\|^2)\, a$ whenever $a \neq 0$, which points in the $-\nabla_x \tilde{h}$ direction, steering the state to satisfy the barrier condition. Equivalently, $u^* = \min(0,\, b/\|a\|^2)\, a$. The regularity assumption $\nabla h(x) \neq 0$ for $x \in \partial\mathcal{C}$ ensures that $a \neq 0$ in a neighborhood of the constraint boundary where intervention is needed. In the interior of $\tilde{\mathcal{C}}(t)$ where $\tilde{h}(x,t) \gg 0$, the constraint is slack and $u^* = 0$. Consequently, the QP is always feasible for any finite noise realization ξ, and infeasible in the degenerate case $\|\xi\| \to \infty$, which occurs with probability zero under Gaussian noise.*

The following result characterizes the distributional shift induced by our guidance.

Theorem 4.2 (Distribution shift for guided sampling) *Let $p(0)$ be the probability distribution generated by the unconstrained sampling process* (1)*, and let $p_{safe}(0)$ be the distribution resulting from the guided dynamics in* (5) *with additive control u. The Kullback-Leibler (KL) divergence between the safe and original distributions is given by:*

$$D_{\mathrm{KL}}(p_{safe}(0) \parallel p(0)) = \frac{1}{2}\mathbb{E}_{p_{safe}(0)}\left[\int_0^T \frac{\|u(x,t)\|^2}{g(t)^2}\mathrm{d}t\right]. \tag{9}$$

We refer the reader to Appendix A.2 for the proof. The minimum-norm control $u^* = \arg\min \|u\|^2$ subject to (7) is a greedy minimizer of the instantaneous contribution to KL divergence between the safe and learned distributions at each sampling instance. While this strategy does not guarantee global optimality over the entire sampling horizon, it provides the tightest instantaneous bound on the distributional shift at each step. The factor $\frac{1}{g(t)^2}$ in (9) reveals that control interventions are cheapest in distributional terms when the noise level $g(t)$ is large. Our constricting safety tube is designed to exploit this structure. Flow-based models perform coarse-to-fine refinement: the high-noise phase ($t \approx T$) establishes global structure, while the low-noise phase ($t \approx 0$) resolves fine details. The safety tube mirrors this progression, maximally relaxed when noise is high and tightening only as the model refines local details. As the tube is loosest precisely when intervention is cheapest, most constraint enforcement is absorbed during the high-noise regime at minimal distributional cost. By the time the model enters the low-noise regime where fine details are resolved, and intervention is expensive, the tube has already guided the trajectory close to the safe set, and $u^* \approx 0$. The model thus retains full authority over the structure and details that determine sample quality. When the safe set $\mathcal{C}$ overlaps significantly with $\mathrm{supp}(p(0))$, the learned drift f_θ already steers samples toward safety, requiring minimal control effort throughout. When $\mathcal{C}$ is disjoint from $\mathrm{supp}(p(0))$, the QP identifies the closest safe distribution under the greedy minimization.

The QP in (8) minimizes $\|u\|^2$ at every state-time pair (x,t) independently. The resulting control u^* is a minimizer of the instantaneous KL divergence rate $\|u\|^2/g(t)^2$. In deterministic regimes $g(t) = 0$, such as flow matching, our framework naturally transitions into a minimal L_2 drift perturbation. Exact minimization of the distribution shift over the full horizon requires solving an optimal control problem over the whole horizon $[0,T]$. This involves the prediction of the behavior of the model, and taking into account future noise realizations. Since the state trajectory under the controlled process depends on the full history of past controls, pointwise minimization does not, in general, yield the global optimal policy over all feasible control sequences. However, u^* in (8) provides a principled strategy that minimizes the instantaneous contribution to the KL divergence between the final distributions at each step, yielding a strong upper bound on the total distributional shift over the whole sampling process. In the stochastic regime when $g(t) > 0$, this greedy policy is the tightest single-step bound available.

Our method of refining the CBF over the sampling horizon stands in contrast to projection-based methods, which apply corrections independently of the noise schedule and therefore pay the same distributional cost at every step, disrupting global structure when applied early and overriding fine details when applied late. The experimental consequence is visible in Figure 4: projection-based enforcement (b) satisfies the constraint but destroys semantic coherence, whereas our CBF-guided sampling (a, c) preserves realistic scene structure by deferring to the model during the critical structure-forming phase. Having established the theoretical foundations for safe sampling, we now turn to practical implementation.

4.2 Discrete-time implementation

We now discuss the discretization of our continuous-time framework into a practical algorithm for guided sampling. We discretize the sampling interval $[0, T]$ into K steps of size $\Delta t = T/K$, with time indices $t_k = T - k\Delta t$ for $k = 0, 1, \ldots, K$, and write $x_k = x(t_k)$ for brevity. The discrete-time evolution under the Euler-Maruyama scheme follows:

$$x_{k-1} = x_k - [f_\theta(x_k, t_k) + u_k(x_k, \xi_k, t_k)]\Delta t + g(t_k)\sqrt{\Delta t}\,\xi_k, \tag{10}$$

where $\xi_k \sim \mathcal{N}(0, I)$. To enforce $x_{k-1} \in \tilde{\mathcal{C}}(t_{k-1})$, we require $\tilde{h}(x_{k-1}, t_{k-1}) \geq 0$. Applying a first-order Taylor expansion of $\tilde{h}$ around (x_k, t_k) and substituting the update (10) yields the linear constraint:

$$\tilde{h}_k + \nabla_x \tilde{h}_k \cdot \Big(-[f_\theta(x_k, t_k) + u_k]\Delta t + g(t_k)\sqrt{\Delta t}\,\xi_k\Big) - \frac{\partial \epsilon}{\partial t}\Delta t \geq (1 - \gamma\Delta t)\,\tilde{h}_k, \tag{11}$$

which is linear in u_k, making the minimum-norm control synthesis a convex QP at each step. The complete procedure is given in Algorithm 1.

We note that the discrete-time constraint (11) is derived via a first-order Taylor series expansion of $\tilde{h}$ around (x_k, t_k), and is therefore only a sufficient condition for $\tilde{h}(x_{k-1}, t_{k-1}) \geq 0$ only up to linearization error. For barrier functions with high curvature or large step sizes Δt, the true nonlinear barrier value could deviate from the linear prediction. In practice, the margin provided by the class-$\mathcal{K}$ term $\gamma(\tilde{h}_k)\Delta t$ and the conservative initialization ϵ_0 absorb this error. Across all experiments in Section 5, we observed no constraint violations at the final sample, empirically validating the linearization for the step sizes used. The discrete-time constraint (11) provides an approximate safety guarantee: the first-order Taylor expansion introduces an $O(\Delta t^2)$ residual that is not formally bounded in our analysis. An additive margin $c > 0$ can be introduced in the initialization ϵ_0 and the class-$\mathcal{K}$ damping γ to provide empirical robustness to this discretization error, as well as provide a margin so that the optimization does not start from the zero-level set for $\tilde{h}(x, t)$. A rigorous discrete-time CBF analysis that establishes explicit step-size-dependent bounds on constraint violation is an important direction for future work. Nonetheless, across all experiments with step sizes ranging from $\Delta t = 0.01$ (Section 5.1) to $\Delta t = 0.005$ (Section 5.3), we observed zero constraint violations at the final sample.

Algorithm 1: CBF-guided safe sampling

Input: Pre-trained model f_θ, noise schedule $g(t)$, safe set $\mathcal{C} = \{x \mid h(x) \geq 0\}$, number of steps K, class-$\mathcal{K}$ function γ

Output: Safe sample $x_0 \in \mathcal{C}$

1. Initialize: Sample $x_K \sim p(T)$ (e.g., $x_K \sim \mathcal{N}(0, I)$), set $t_K = T$, $\Delta t = T/K$

2. for $k = K, K-1, \ldots, 1$ **do**

- **3.** Sample noise: $\xi_k \sim \mathcal{N}(0, I)$
- **4.** Compute noise-induced drift: $d_{\text{noise}} = g(t_k)\sqrt{\Delta t}\,\xi_k$
- **5.** Compute constricting barrier: $\tilde{h}_k = h(x_k) + \epsilon(x_K, t_k)$
- **6.** Solve safety-constrained QP:

$$\begin{aligned} \min_{u_k} \quad & \frac{1}{2}\|u_k\|^2 \\ \text{s.t.} \quad & \tilde{h}_k + \nabla_x \tilde{h}_k \cdot (-[f_\theta(x_k, t_k) + u_k]\Delta t + d_{\text{noise}}) - \frac{\partial \epsilon}{\partial t}(x_K, t_k)\Delta t \geq (1 - \gamma(\tilde{h}_k)\Delta t)\,\tilde{h}_k \end{aligned}$$

- **7.** Apply control and update state: $x_{k-1} = x_k - f_\theta(x_k, t_k)\Delta t - u_k^*\Delta t + d_{\text{noise}}$
- **8.** Update time: $t_{k-1} = t_k - \Delta t$

9. return x_0

5 Experiments

We validate our framework across three domains that demonstrate its versatility and practical effectiveness. In each experiment, we apply our CBF-guided sampling to pre-trained, off-the-shelf generative models without any retraining or architectural modifications, demonstrating the modularity claimed in Contribution 2. For each experiment, our goal is to demonstrate that the safe sampling guarantees of Theorem 4.1 hold empirically across qualitatively distinct constraint types.

All experiments use the linear constriction $\epsilon(x(T), t) = \epsilon_0 \cdot \frac{t}{T}$ where $\epsilon_0 = \max(0, -h(x(T))) + c$, where $c > 0$ is a small positive margin. This margin prevents the barrier from operating exactly on its zero level set at initialization, where numerical precision could cause spurious violations. In all experiments, we set $c = 0.1$. We choose the class-$\mathcal{K}$ function $\gamma(h) = \alpha h$ with $\alpha = 0.5$. The QPs in Algorithm 1 are solved using CVXPY with the OSQP solver. Our implementation is available at [anonymized for review]. Our experiments were conducted on an Intel i9-9900 machine with 128GB RAM and an Nvidia Quadro RTX 4000 GPU.

5.1 Physics-consistent trajectory generation for the Lorenz system

In this experiment, we demonstrate the behavior of our guidance scheme on a synthetic testbed. We seek to generate trajectories of the Lorenz system with dynamics,

$$\begin{aligned} \dot{z}_1 &= \sigma(z_2 - z_1) \\ \dot{z}_2 &= z_1(\rho - z_3) - z_2 \\ \dot{z}_3 &= z_1 z_2 - \beta z_3 \end{aligned} \tag{12}$$

where we set $\sigma = 10, \rho = 28, \beta = \frac{8}{3}$. This system exhibits a characteristic behavior called the Lorenz attractor, famously known as the butterfly effect, where it is highly sensitive to the initial condition. Our goal is to ensure that sampled trajectories satisfy the true physics encoded in (12). Given an initial condition $z(0) = [z_1(0)\ z_2(0)\ z_3(0)]^\top$, we seek to sample an entire trajectory up to 10 seconds of evolution. We use a physical time discretization $\Delta_l = 0.01$, which results in $L = 1000$ sampling steps. For the sake of brevity, we denote the state sampled at time $l\Delta_l$ as $z(l\Delta_l) = z^l$. We sample a full trajectory $x(0) = [{z^0}^\top\ {z^1}^\top\ \cdots\ {z^L}^\top]^\top$. Note that l is the sampling step in physical time for the Lorenz system, whereas we reserve k for the sampling time for the diffusion model.

We select the Lorenz oscillator because its ground-truth solution is available via numerical integration, providing an exact reference for validating our guidance scheme. In practice, our framework targets settings where the governing equations are high-dimensional or computationally expensive to integrate, such as turbulent flows or multi-scale PDEs, and where a generative model serves as an efficient surrogate. Unlike traditional numerical methods that accumulate integration error over time, our approach generates the entire trajectory as a single high-dimensional sample. The Lorenz system thus serves as a controlled testbed to verify that our CBF-guided sampling can recover underlying physical laws and maintain temporal consistency across the entire sequence before deployment in more complex domains.

We train a diffusion model with synthetic data generated by numerically integrating (12) from initial conditions uniformly sampled from the region $[-2, 2]^3$. The model is implemented as a conditional DDPM with a U-Net architecture featuring 4 downsampling blocks, each with 64, 128, 256, and 512 channels respectively, trained for 1000 epochs with a cosine noise schedule. As we seek adherence to the true dynamics (12), we define the barrier function as:

$$h(x) = e - \frac{1}{L} \sum_{l=0}^{L-1} \left\| \frac{z^{l+1} - z^l}{\Delta_l} - \dot{z}^l \right\|^2 \tag{13}$$

where $\dot{z}^l$ is given by the true vector field (12), and e is an error tolerance in the average physics adherence. The safe set $\mathcal{C} = \{x \mid h(x) \geq 0\}$ enforces the physics up to a tolerance $e = 0.001$.

Figure 2 presents key results of our CBF-guided sampling for Lorenz system trajectory generation. Figure 2(a) shows trajectories in the phase plane. The true ODE solution (dashed black) is closely tracked by

our CBF-guided diffusion model (blue), while unconstrained sampling (red) produces trajectories that deviate significantly from the true physics. The unconstrained sample still produces the characteristic butterfly effect, but it is not the true trajectory that is followed by the system. This arises due to the fact that the diffusion model has learned to produce samples that exhibit the butterfly attractor by seeing ground truth data in training. This highlights a subtle but critical failure mode: a generative model can produce samples that are statistically indistinguishable from real data yet physically incorrect, and any downstream task that consumes such samples, whether a controller, a simulator, or a decision-making system, would operate on corrupted inputs without any indication of the violation.

Figure 2(b) reveals the evolution of the safety tube. At $t = 1$, the noise sample $x(T) \sim \mathcal{N}(0, I)$ has large cumulative physics error, requiring relaxation $\epsilon_0 \approx 16$—over four orders of magnitude larger than our target tolerance $e = 0.001$. The blue line shows the linear constriction $\epsilon(x(T), t) = \epsilon_0 \cdot (t/T)$ to zero at $t = 0$. The constriction barrier $\tilde{h}(x, t)$ (red) remains non-negative throughout, verifying reverse invariance (Theorem 4.1). Fluctuations during $t \approx 0.8$-1.0 reflect the high-noise regime where $g(t)$ is large. At $t = 0$, $\tilde{h}(x, 0) \approx 0$ with $\epsilon(x(T), 0) = 0$ implies $h(x(0)) \approx 0$, confirming that our minimum-norm control maximally exploits the constraint tolerance. Figure 2(c) shows control effort $\|u\|^2$ concentrated at sampling onset ($t \approx 0.8$-1.0), peaking around 130-140, then rapidly decaying to near-zero by $t \approx 0.6$. This front-loaded intervention exploits that control is cheaper when $g(t)$ is large. The minimal effort for $t < 0.6$ indicates that the pre-trained model has implicitly learned the Lorenz system structure from training data, requiring only minor corrections to enforce hard guarantees.

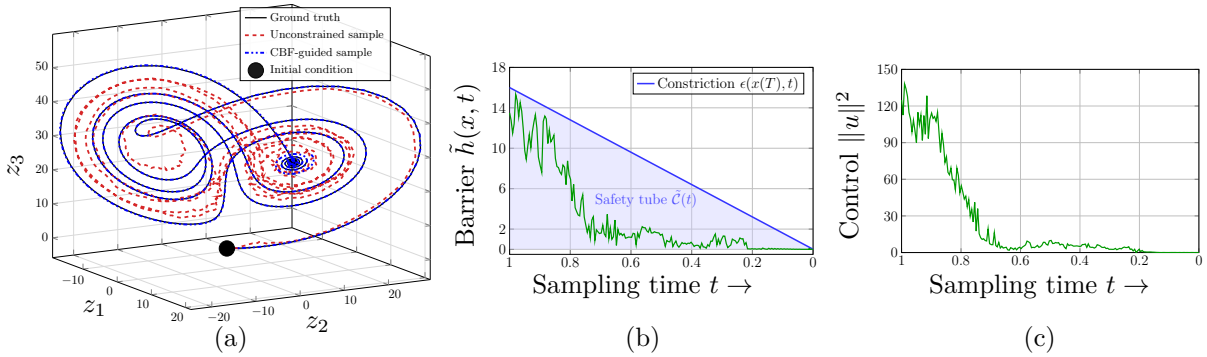


Figure 2: Lorenz system trajectory generation with CBF guidance. (a) Phase portrait showing the true ODE solution (dashed black). Our CBF-guided diffusion model (blue) closely tracks the true dynamics, while unconstrained sampling (red) produces physically inconsistent trajectories. (b) Evolution of the constricting barrier $\tilde{h}(x, t)$ (green) and relaxation $\epsilon(x(T), t)$ (blue) during sampling. The constriction goes from $\epsilon_0 \approx 16$ to 0, accommodating the large initial physics violation. The barrier remains non-negative throughout, verifying reverse invariance. (c) Control effort $\|u\|^2$ over sampling time, concentrated at the onset of sampling when the initial noise sample must be corrected, then rapidly diminishing as the trajectory becomes physics-consistent.

5.2 Constrained image generation

We demonstrate our framework on image synthesis tasks using the off-the-shelf DDPM-bedroom-256 [4] model from Hugging Face Diffusers, trained on the LSUN bedroom dataset (Yu et al., 2015). Images are represented as $x \in \mathbb{R}^{256\times256\times3}$ with RGB values normalized to $[-1, 1]$.

A key feature of our framework in this setting is that we define one barrier function per constrained pixel, rather than a single aggregate barrier over all pixels. For each pixel $\mathbf{p}$ in the constrained region, we define an independent barrier $h_{\mathbf{p}}(x)$ and enforce $h_{\mathbf{p}}(x) \geq 0$ separately. The resulting QP at each sampling step contains $|R|$ linear constraints, where R is the set of constrained pixels. Crucially, since each per-pixel barrier $h_{\mathbf{p}}$ depends only on the three RGB values $x_{\mathbf{p}} \in \mathbb{R}^3$ at pixel $\mathbf{p}$. This sparsity implies that the multi-constraint QP over the full image space $\mathbb{R}^{256\times256\times3}$ decomposes into $|\mathcal{R}|$ independent three-dimensional QPs, one per pixel.

[4] `https://huggingface.co/google/ddpm-bedroom-256`

Each sub-problem admits the closed-form solution from Remark 2, enabling efficient parallel computation. In our experiments, we solve the full QP using OSQP via CVXPY, which exploits this sparsity internally.

5.2.1 Location and content constraints

In this experiment, we enforce specific visual content at designated pixel locations, while retaining semantic meaning with the rest of the image. Given any reference image, we constrain a rectangular region $R \subseteq \mathbb{N}^2$ of dimension $h \times w$ in the generated image $x \in \mathbb{R}^{256\times256\times3}$ to match the reference.

To ensure the reference image appears in the prescribed location, we enforce pixel-level constraints inside the region R. Further, we modulate the constraint strength near boundary of R so that the diffusion model can smoothly fill the edges of the image with semantic information. We define a spatially-varying mask function $v : \mathbb{N}^2 \to [0, 1]$ for each pixel $\mathbf{p} \in \mathbb{R}^{256\times256\times3}$:

$$v(\mathbf{p}) = \begin{cases} 1 & \text{if } \mathbf{p} \in \text{interior}(R) \\ \text{smooth decay to } 0 & \text{if } \mathbf{p} \in \text{boundary}(R) \\ 0 & \text{if } \mathbf{p} \notin R \end{cases}$$

where the boundary region is defined as pixels within 5% of the edge of R, and the decay is implemented via a linear ramp. Let $x^*_{\mathbf{p}} \in \mathbb{R}^3$ denote the RGB values of the reference image $\mathbf{P}$ at the corresponding position within R. For each pixel $\mathbf{p} \in R$, we define the barrier function:

$$h_{\mathbf{p}}(x) = e - v(\mathbf{p}) \cdot \|x_{\mathbf{p}} - x^*_{\mathbf{p}}\|^2, \tag{14}$$

where $e = 0.005$ is the error tolerance and $x_{\mathbf{p}} \in \mathbb{R}^3$ denotes the RGB pixel values of x at location $\mathbf{p}$. The safe set is the intersection $\mathcal{C} = \bigcap_{\mathbf{p}\in R}\{x \mid h_{\mathbf{p}}(x) \geq 0\}$, which enforces pixel-level fidelity in the interior of R where $v(\mathbf{p}) = 1$, while allowing deviations near boundaries where $v(\mathbf{p}) < 1$. The smooth decay $v(\mathbf{p}) \to 0$ allows the diffusion model to alter the edges of the reference image $\mathbf{P}$ as required for natural blending. Each pixel has its own constricting barrier $\tilde{h}_{\mathbf{p}}(x, t) = h_{\mathbf{p}}(x) + \epsilon_{\mathbf{p}}(x(T), t)$ with a linear relaxation $\epsilon_{\mathbf{p}}(x(T), t) = \epsilon_{0,\mathbf{p}} \cdot t/T$, where $\epsilon_{0,\mathbf{p}} = \max(0, -h_{\mathbf{p}}(x(T))) + c$, where $c = 0.01$ is a small positive margin.

Figure 3 demonstrates constrained image generation with a 50×70 pixel window region placed at position $(40, 150)$. Both generated images (b-c) preserve the reference window (a) exactly, validating Theorem 4.1. With 200 sampling steps (c), the model generates a coherent bedroom scene with semantically appropriate context—bed, lamps, and furniture properly scaled and lit relative to the window. The spatially-varying mask enables smooth blending at boundaries without visible artifacts. With only 50 sampling steps (b), visual quality degrades in unconstrained regions, yet the window remains perfectly preserved. This demonstrates that the CBF shield guarantees constraint satisfaction regardless of sampling duration, while unconstrained regions follow standard diffusion model behavior. This arises due to how the constriction ϵ is constructed with conditions in Definition 1.

This experiment demonstrates the modularity of our framework. We apply our method to an off-the-shelf pre-trained model without any architectural modifications or retraining, enforcing hard spatial constraints that would be impossible to guarantee with soft guidance methods. Additional results with different reference images and placements are provided in Appendix B.1.

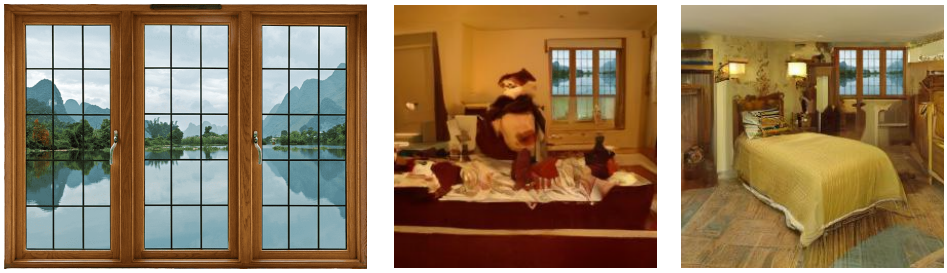
5.2.2 Regional color intensity constraints

We demonstrate enforcement of regional color constraints by constraining the lower one-third of the image to maintain specified color intensities. Define the constrained region as:

$$R_{\text{lower}} = \{(i, j) \in \mathbb{N}^2 : 171 \leq i \leq 256,\ 1 \leq j \leq 256\},$$

which corresponds to the lower one-third of the image. The target color intensity is $x^*_{\mathbf{p}} \in [-1, 1]^3$ for all $\mathbf{p} \in R_{\text{lower}}$. To allow natural blending near the boundary, we employ a spatially varying mask function $v : \mathbb{N}^2 \to [0, 1]$ that modulates constraint strength:

$$v(\mathbf{p}) = \begin{cases} \frac{i-i_{\min}}{i_{\max}-i_{\min}} \cdot v_{\max} + \left(1 - \frac{i-i_{\min}}{i_{\max}-i_{\min}}\right) \cdot v_{\min}, & \text{if } \mathbf{p} = (i, j) \in R_{\text{lower}}, \\ 0, & \text{otherwise}, \end{cases} \tag{15}$$



(a) (b) (c)

Figure 3: Spatially constrained image generation with DDPM-bedroom-256. (a) Reference window patch (50×70 pixels). (b) Generated image with 50 sampling steps. (c) Generated image with 200 sampling steps. Both (b) and (c) preserve the reference window exactly at the specified location while generating bedroom context. Higher sampling steps improve generation quality in unconstrained regions, but constraint satisfaction is guaranteed in both cases.

where $i_{\min} = 171$, $i_{\max} = 256$ denote the row boundaries of R_{lower}, and $(v_{\min}, v_{\max}) \in [0, 1]^2$ the constraint strength, respectively. For each pixel $\mathbf{p} \in R_{\text{lower}}$, the barrier function is:

$$h_{\mathbf{p}}(x) = e - v(\mathbf{p}) \cdot \|x_{\mathbf{p}} - x^*_{\mathbf{p}}\|^2, \tag{16}$$

where $e = 0.05$ is the error tolerance, and $x_{\mathbf{p}} \in \mathbb{R}^3$ denotes the RGB pixel values of x at pixel $\mathbf{p}$. The safe set $C = \{x \in \mathbb{R}^{256\times256\times3} \mid h(x) \geq 0\}$ represents images with pixel-level color fidelity weighted by the mask function $v(\mathbf{p})$, where higher mask values enforce stricter adherence to the target intensity. We use the linear constriction scheme $\epsilon(x(T), t) = \epsilon_0 \cdot t/T + c$ where $\epsilon_0 = \max(0, -h(x(T)))$.

In this experiment, we use different mask configurations and color intensities and compare our constricting CBF guidance scheme with projection-based constraint enforcement in Zampini et al. (2025). This experiment illustrates how our coarse-to-fine constraint enforcement retains semantic meaning with the rest of the image whereas projection schemes can lose semantic meaning:

1. *Moderate constraint* in Figure 4(a): The target pixel color is black, $x^*_{\mathbf{p}} = (-0.9, -0.9, -0.9)$, and the spatial mask $v(\mathbf{p})$ varies from $v_{\max} = 0.5$ at the bottom to $v_{\min} = 0$ at the upper boundary of R_{lower}. This allows the diffusion model freedom to generate detailed, realistic textures while satisfying the per-pixel color constraints.

2. *Projection* in Figure 4(b): We implement the projection-based constraint enforcement scheme in Zampini et al. (2025) and compare the visual quality. The constraint is the same as in the earlier case where the bottom-third of the image needs to be black $x^*_{\mathbf{p}} = (-0.9, -0.9, -0.9)$, and the spatial mask $v(\mathbf{p})$ varies from $v_{\max} = 0.5$ at the bottom to $v_{\min} = 0$ at the upper boundary of R_{lower}. Due to projection at each step, we get a black-tape effect, which achieves constraint enforcement but semantic meaning is lost.

3. *Weak constraint* in Figure 4(c): The target pixel intensity is $x_{\mathbf{p}} = (-0.4, -0.3, -0.3)$, corresponding to a light brown color. The mask values range from $(v_{\min}, v_{\max}) = (0, 0.2)$, applying weak constraints. This allows the diffusion model to almost freely generate an image, relying on the guidance to only nudge individual pixels toward the target color.

All images successfully satisfy the color constraint in the lower region while maintaining semantic coherence with realistic bedroom layouts. However, the choice of mask strength directly affects the balance between strict constraint adherence and natural visual appearance, with lower mask values allowing greater model freedom to generate detailed, realistic textures within the constrained color range. Figure 4 illustrates how our CBF-guided sampling framework enables precise control over the strength and extent of constraint enforcement. The spatially-varying mask function $v(p)$ provides a smooth control between strict enforcement

and model freedom, allowing practitioners to balance safety enforcement with perceptual quality based on application requirements. Importantly, all three configurations achieve 100% constraint satisfaction. No generated image violates $h(x) \geq 0$, demonstrating the formal safety guarantees of Theorem 4.1 even under weak constraint settings.

(a) (b) (c)

Figure 4: Regional color intensity constraints. Images are enforced with color values in the lower third. (a) *Moderate intensity:* The target pixel color is black, and the spatial mask $v(\mathbf{p})$ goes from $0.5 \to 0$, allowing the diffusion model to produce clean furniture. (b) *Projection:* Projection-based constraint enforcement (Zampini et al., 2025), constraints are the same as in (a), with the target pixel color black in the bottom one-third of the image. Projection causes intense constraint enforcement, which leads to loss of semantic information. (c) *Low intensity:* Target color is brown, and the spatial mask $v(\mathbf{p})$ goes from 0.2 at the bottom to 0 at the lower-third boundary, producing a semantically meaningful image of a room with a brown carpet.

5.3 Smooth robot policy generation

We demonstrate our framework on robotic manipulation using the pre-trained Diffusion Policy model (Chi et al., 2025) for the Push-T task. The Push-T task requires a planar robotic arm to push a T-shaped block to a target pose, as depicted in Figure 5(a). The diffusion policy generates a sequence of actions $x(0) = \{a_0, a_1, \ldots, a_S\}$, where $S = 15$ is the action horizon and each action $a_s \in \mathbb{R}^2$ represents an end-effector velocity command. We seek safe samples of action chunks $x(0)$ by applying Algorithm 1 during the diffusion sampling process, without any retraining or modification to the model architecture. On physical robotic hardware, abrupt changes in velocity commands pose concrete safety risks. Large accelerations violate actuator rate limits, inducing torque spikes that can damage motors and gearboxes. In contact-rich tasks such as Push-T, jerky motions destabilize the grasp or push contact, leading to task failure or uncontrolled object motion. Smoothness of the commanded action sequence is therefore not merely a quality metric but an operational safety requirement for real-world deployment. We encode this requirement as a hard constraint by bounding the average control variation, or the jerk, across the action horizon. Specifically, we define the barrier function as:

$$h(x) = e_{\text{smooth}} - \frac{1}{S} \sum_{s=0}^{S-1} \frac{\|a_{s+1} - a_s\|^2}{\Delta_s}, \tag{17}$$

where $e_{\text{smooth}} = 1.5$ is the tolerance that bounds the maximum allowed jerk magnitude, and Δ_s is the time discretization of the environment. The safe set $\mathcal{C} = \{x \mid h(x) \geq 0\}$ encodes the set of all action sequences whose average control variation magnitude remains below the prescribed threshold e_{smooth}. We apply our Algorithm 1 with the same linear constriction scheme as in the previous experiments.

Table 1 presents quantitative results averaged over 100 episodes. We compare our CBF-guided sampling against the original Diffusion Policy with DDPM sampling (DP) and DDIM sampling (DP-DDIM), using the pre-trained checkpoints from the authors' repository[5]. Our CBF-guided sampling achieves a mean reward

[5] `https://github.com/real-stanford/diffusion_policy`

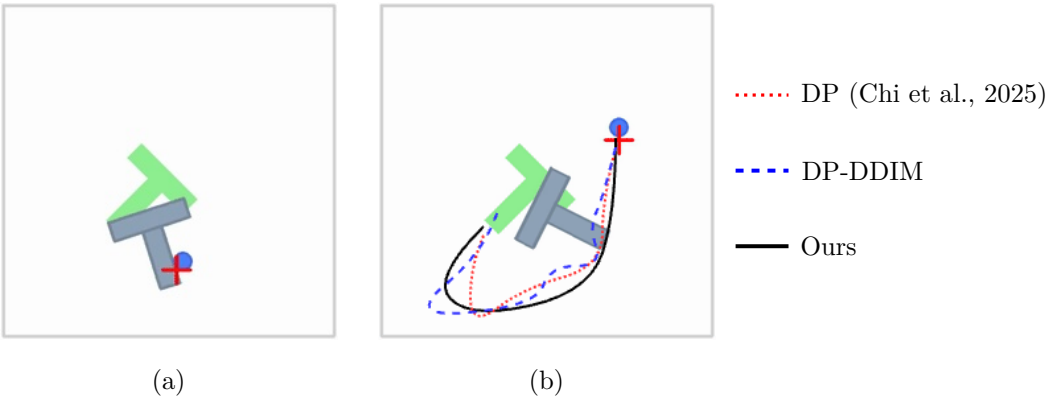


Figure 5: Smooth action generation for Push-T robotic manipulation. (a) The Push-T environment: a planar robot arm must push the T-shaped block (gray) to the target pose (green). (b) End-effector trajectories for a representative episode. Unconstrained Diffusion Policy (DP, dotted red) exhibits sharp directional changes during pushing. DP-DDIM (dashed blue) produces more erratic motion due to fewer sampling steps. Our CBF-guided sampling (solid black) generates a smooth trajectory that satisfies the acceleration constraint at every step while achieving the same task reward.

Table 1: Comparison on Push-T with smoothness constraints over 100 episodes

Metric	**DP**	**DP-DDIM**	**CBF-guided**
Mean reward	0.92	0.90	0.92
Smoothness violation	12	16	**0**
Sampling steps K	100	10	100
Mean inference time (ms)	47.05	4.57	62.92

of 0.92, matching the original Diffusion Policy, while guaranteeing zero smoothness violations across all episodes. In contrast, unconstrained DP and DP-DDIM exceed the smoothness threshold ϵ_{smooth} an average of 12 and 16 times per episode, respectively. This demonstrates that the learned policy does not inherently satisfy smoothness requirements despite being trained on smooth demonstrations. The diffusion sampling process itself introduces high-frequency artifacts that violate the constraint. Figure 5(b) visualizes the end-effector trajectory for a representative episode. The unconstrained DP trajectory (dotted red) exhibits sharp directional changes, particularly during the pushing phase. DP-DDIM (dashed blue), which uses fewer sampling steps, produces even more erratic motion. Our CBF-guided trajectory (black) follows a visibly smoother path while achieving the same task objective. The smoothness is enforced throughout the diffusion sampling process rather than applied as post-hoc filtering, ensuring that every generated action chunk satisfies the constraint by construction. The additional computational cost of solving the QP at each sampling step is modest. Inference time increases from 47 ms to 63 ms per sample, a 34% overhead that remains well within real-time requirements for the 10 Hz control loop of the Push-T environment. DP-DDIM achieves significantly faster inference (4.6 ms) by using only 10 sampling steps, but at the cost of higher smoothness violations and slightly lower task reward.

This experiment validates our framework in a robotic planning setting where the generative model operates over structured action sequences rather than images or physical trajectories. The constraint is enforced purely in action space, requiring no forward dynamics model or state prediction. Extending to state-space safety constraints, such as collision avoidance, would require access to a dynamics model (analytical or learned) to propagate the effect of actions to states. We discuss this as a natural extension in Section 6.

6 Limitations and future work

Our framework assumes access to a continuously differentiable barrier function $h(x)$ that precisely encodes the safety specification. When the notion of safety is well-defined and analytically expressible, such as physics residuals (Section 5.1), pixel-level constraints (Section 5.2), or action smoothness bounds (Section 5.3), our method provides deterministic guarantees. However, in domains where safety is ambiguous or difficult to formalize, such as filtering semantically inappropriate content in images, constructing a suitable barrier is nontrivial. In our preliminary experiments, we attempted to use CLIP-based confidence scores and neural network classifiers as barrier functions for semantic constraints. These attempts were unsuccessful on several occasions. The classifiers occasionally assigned high confidence to unsafe samples, causing the CBF condition to be trivially satisfied and the safety filter to remain inactive. This failure mode is well-documented in the adversarial robustness literature and stems from the fact that learned classifiers extract features that do not reliably align with the safety boundary. When the barrier function itself is unreliable, obtaining formal guarantees on safety with respect to the intended safety semantics is unachievable.

Several extensions follow naturally from our framework. First, extending the approach to latent diffusion models would enable scalability to higher-resolution generation. We investigated extending our framework to latent diffusion models (Rombach et al., 2022) where the diffusion process operates in a compressed latent space $z = E(x)$ and the constraint is defined in image space. While the QP structure is preserved, the latent gradient is computed via a single backward pass through the decoder. Our preliminary experiments with Stable Diffusion v1.5 revealed that exact constraint satisfaction does not transfer from latent to image space (see Appendix B.2). The VAE decoder is not a diffeomorphism: it is locally non-invertible and introduces reconstruction artifacts that distort the constraint boundary. As a result, the CBF shield successfully biases the latent trajectory toward the constrained region but cannot guarantee pixel-level fidelity in the decoded image. Achieving exact constraints in latent diffusion models likely requires either (i) a constraint-aware fine-tuning of the decoder to improve local invertibility near the constraint boundary, or (ii) a hybrid approach that applies coarse guidance in latent space and a final correction step in pixel space after decoding. Second, enforcing state-space safety constraints in robotic policies, such as collision avoidance, requires a dynamics model to map actions to states. Combining our framework with learned dynamics models or neural ODEs is a promising direction. Third, our current implementation solves one QP per sampling step using a general-purpose solver. Using an MPC framework across many steps could significantly yield tighter bounds on the KL divergence over the entire sampling horizon. In particular, a receding-horizon formulation that anticipates the constriction schedule over multiple future steps could yield control policies that are globally optimal, rather than greedy at each step, potentially tightening the bound over the learned sampling process.

7 Conclusion

We introduced a framework for enforcing hard constraints on flow-based generative models by framing guided sampling as a problem of control synthesis. Our approach leverages constricting Control Barrier Functions (CBFs) to define a safety tube that is relaxed at the initial noise distribution and progressively constricts to the target safe set. The resulting minimum-norm QP synthesizes a feedback control input that provably maintains samples within the safety tube (Theorem 4.1) while minimizing the instantaneous contribution to the distributional shift from the learned model, as quantified by the KL divergence (Theorem 4.2). We validated the modularity and effectiveness of this framework across physically-consistent trajectory generation, constrained image synthesis, and smooth action generation for robotic manipulation, using off-the-shelf pre-trained models. Our framework maintains semantic fidelity of the model while ensuring constraint adherence. As generative models are increasingly deployed in safety-critical systems, this framework provides a deterministic safety layer that complements the expressiveness of generative sampling models.

References

A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada. Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, 62(8):3861–3876, 2016.

A. D. Ames, S. Coogan, M. Egerstedt, G. Notomista, K. Sreenath, and P. Tabuada. Control barrier functions: Theory and applications. In *2019 18th European Control Conference (ECC)*, pp. 3420–3431, 2019.

F. Blanchini and S. Miani. *Set-theoretic methods in control*, volume 78. Springer, 2008.

C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

A. Clark. Control barrier functions for stochastic systems. *Automatica*, 130:109688, 2021.

X. Dai, Z. Yang, D. Yu, S. Zhang, H. Sadeghian, S. Haddadin, and S. Hirche. Safe flow matching: Robot motion planning with control barrier functions. *arXiv preprint arXiv:2504.08661*, 2025.

P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

N. Fishman, L. Klarner, V. De Bortoli, E. Mathieu, and M. J. Hutchinson. Diffusion models for constrained domains. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=xuWTFQ4VGO`.

D. Gadginmath and F. Pasqualetti. Dynamics-aware diffusion models for planning and control. In *64th IEEE Conference on Decision and Control (CDC)*, 2025.

Y. Guo, H. Yuan, Y. Yang, M. Chen, and M. Wang. Gradient guidance for diffusion models: An optimization perspective. *Advances in Neural Information Processing Systems*, 37:90736–90770, 2024.

J. Ho and T. Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL `https://openreview.net/forum?id=qw8AKxfYbI`.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.

M. Janner, Y. Du, J. Tenenbaum, and S. Levine. Planning with diffusion for flexible behavior synthesis. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9902–9915. PMLR, 17–23 Jul 2022.

Z. Li, D. Chen, M. Fan, C. Chen, Y. Li, Y. Wang, and W. Zhou. Responsible diffusion models via constraining text embeddings within safe regions. In *Proceedings of the ACM on Web Conference 2025*, pp. 1588–1601. Association for Computing Machinery, 2025.

B. Liao, S. Chen, H. Yin, B. Jiang, C. Wang, S. Yan, X. Zhang, X. Li, Y. Zhang, Q. Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12037–12047, 2025.

Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and B. Poole. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=PqvMRDCJT9t`.

A. Lou and S. Ermon. Reflected diffusion models. In *International Conference on Machine Learning*, pp. 22675–22701. PMLR, 2023.

K. Mizuta and K. Leung. Cobl-diffusion: Diffusion-based conditional robot planning in dynamic environments using control barrier and lyapunov functions. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 13801–13808. IEEE, 2024.

B. Øksendal. Stochastic differential equations. In *Stochastic differential equations: an introduction with applications*, pp. 38–50. Springer, 2003.

G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

S. Prajna, A. Jadbabaie, and G. J. Pappas. Stochastic safety verification using barrier certificates. In *2004 43rd IEEE conference on decision and control (CDC)*, volume 1, pp. 929–934. IEEE, 2004.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

R. Römer, L. Brunke, M. Schuck, and A. P. Schoellig. Safe offline reinforcement learning using trajectory-level diffusion models. In *ICRA 2024 Workshop - Back to the Future: Robot Learning Going Probabilistic*, 2024.

P. Schramowski, M. Brack, B. Deiseroth, and K. Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22522–22531, 2023.

O. So, A. Clark, and C. Fan. Almost-sure safety guarantees of stochastic zero-control barrier functions do not hold. *arXiv preprint arXiv:2312.02430*, 2023.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

U. Utkarsh, P. Cai, A. Edelman, R. Gomez-Bombarelli, and C. V. Rackauckas. Physics-constrained flow matching: Sampling generative models with hard constraints. *arXiv preprint arXiv:2506.04171*, 2025.

T. Weiss, E. Mayo Yanes, S. Chakraborty, L. Cosmo, A. M. Bronstein, and R. Gershoni-Poranne. Guided diffusion for inverse molecular design. *Nature Computational Science*, 3(10):873–882, 2023.

W. Xiao, T.-H. Wang, C. Gan, R. Hasani, M. Lechner, and D. Rus. Safediffuser: Safe planning with diffusion probabilistic models. In *The Thirteenth International Conference on Learning Representations*, 2023.

J. Yang, S. Jang, and S. Han. Safeflowmatcher: Safe and fast planning using flow matching with control barrier functions. *arXiv preprint arXiv:2509.24243*, 2025.

F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

H. Yuan, K. Huang, C. Ni, M. Chen, and M. Wang. Reward-directed conditional diffusion: Provable distribution estimation and reward improvement. *Advances in Neural Information Processing Systems*, 36:60599–60635, 2023.

S. Zampini, J. K. Christopher, L. Oneto, D. Anguita, and F. Fioretto. Training-free constrained generation with stable diffusion models. *arXiv preprint arXiv:2502.05625*, 2025.

A Appendix: Proofs for Theorem 4.1 and Theorem 4.2

A.1 Proof of Theorem 4.1 (Reverse invariance of guided sampling process)

We first establish initial feasibility at the onset of the sampling process at $t = T$. Given $x(T)$, an arbitrary sample drawn from the noise distribution $p(T)$, the constricting barrier function is:

$$\tilde{h}(x(T), T) = h(x(T)) + \epsilon(x(T), T).$$

By the condition $\epsilon(x(T), T) \geq -h(x(T))$, it follows immediately that $\tilde{h}(x(T), T) \geq 0$, which implies $x(T) \in \tilde{\mathcal{C}}(T)$. Next, we consider the evolution of the state $x(t)$ under the controlled dynamics (5). The

key observation is that in our setting, we have full knowledge of the noise process: at each instant t, the noise realization $\xi(t)$ is observed *before* the control $u(x, \xi, t)$ is synthesized. Consequently, the result follows if we can prove reverse invariance for each realization of the noise path $\{\xi(t)\}_{t\in[0,T]}$, where the controlled dynamics reduce to the following deterministic time-varying ordinary differential equation

$$\dot{x}(t) = \tilde{f}(x, t) := f_\theta(x, t) + u(x, \xi(t), t) + g(t)\xi(t), \tag{18}$$

For each such realization, we verify the conditions of the generalization of Nagumo's theorem (Blanchini & Miani, 2008, Section 4.2.2) for the time-varying set $\tilde{\mathcal{C}}(t) = \{x \mid \tilde{h}(x, t) \geq 0\}$. Existence and uniqueness of solutions to (18) for each realization of the noise process ξ hold by (Øksendal, 2003, Theorem 5.2.1) By condition (7), for any (x, t) with $\tilde{h}(x, t) = 0$, the control $u(x, \xi(t), t)$ ensures

$$\nabla\tilde{h}(x, t) \cdot \tilde{f}(x, t) + \frac{\partial\tilde{h}(x, t)}{\partial t} \leq \gamma(\tilde{h}(x, t)) = 0, \tag{19}$$

where the equality $\gamma(\tilde{h}(x, t)) = 0$ follows from γ being a class-$\mathcal{K}$ function evaluated at zero. Since (18) is a deterministic ODE for each fixed noise path, the generalization of Nagumo's theorem applies directly, yielding reverse invariance of $\tilde{\mathcal{C}}(t)$: if $x(T) \in \tilde{\mathcal{C}}(T)$, then $x(t) \in \tilde{\mathcal{C}}(t)$ for all $t \in [0, T]$.

Since this argument holds for every noise realization $\{\xi(t)\}_{t\in[0,T]}$, we conclude that $\tilde{h}(x(t), t) \geq 0$ for all $t \in [0, T]$. At $t = 0$, we have $\tilde{h}(x(0), 0) = h(x(0)) + \epsilon(x(T), 0) = h(x(0)) \geq 0$, so $x(0) \in \mathcal{C}$. □

A.2 Proof of Theorem 4.2 (Distributional shift wrt to guidance)

Let P denote the probability measure associated with the unconstrained process governed by the drift f_θ, and Q denote the measure associated with the controlled process governed by $f_\theta + u$. We assume $g(t) > 0$ for all $t \in [0, T]$. Recall that the sampling process runs in reverse time from T to 0. To apply standard results, we reparametrize the process in forward time via $\tau = T - t$, so that τ increases from 0 to T as sampling progresses. Under this substitution, the unconstrained and controlled SDEs become

$$\mathrm{d}x = \bar{f}_\theta(x, \tau)\,\mathrm{d}\tau + \bar{g}(\tau)\,\mathrm{d}W_\tau, \qquad \mathrm{d}x = [\bar{f}_\theta(x, \tau) + \bar{u}(x, \tau)]\,\mathrm{d}\tau + \bar{g}(\tau)\,\mathrm{d}W_\tau, \tag{20}$$

where $\bar{f}_\theta(x, \tau) = -f_\theta(x, T - \tau)$, $\bar{g}(\tau) = g(T - \tau)$, $\bar{u}(x, \tau) = -u(x, T - \tau)$, and W_τ is a standard forward-time Wiener process. By Girsanov's Theorem, the Radon-Nikodym derivative of the controlled measure Q with respect to the unconstrained measure P over $[0, T]$ is

$$\frac{\mathrm{d}Q}{\mathrm{d}P} = \exp\left(\int_0^T \frac{\bar{u}(x_\tau, \tau)}{\bar{g}(\tau)} \cdot \mathrm{d}W_\tau - \frac{1}{2}\int_0^T \frac{\|\bar{u}(x_\tau, \tau)\|^2}{\bar{g}(\tau)^2}\,\mathrm{d}\tau\right). \tag{21}$$

The KL divergence is the expected log-likelihood ratio under Q:

$$D_{\mathrm{KL}}(Q \,\|\, P) = \mathbb{E}_Q\left[\ln\frac{\mathrm{d}Q}{\mathrm{d}P}\right] = \mathbb{E}_Q\left[\int_0^T \frac{\bar{u}}{\bar{g}} \cdot \mathrm{d}W_\tau - \frac{1}{2}\int_0^T \frac{\|\bar{u}\|^2}{\bar{g}^2}\,\mathrm{d}\tau\right]. \tag{22}$$

By Girsanov's Theorem, the process $W_\tau^Q = W_\tau - \int_0^\tau \frac{\bar{u}}{\bar{g}}\,\mathrm{d}s$ is a Wiener process under Q. Substituting $\mathrm{d}W_\tau = \mathrm{d}W_\tau^Q + \frac{\bar{u}}{\bar{g}}\,\mathrm{d}\tau$ into (22) yields

$$\begin{aligned} D_{\mathrm{KL}}(Q \,\|\, P) &= \mathbb{E}_Q\left[\int_0^T \frac{\bar{u}}{\bar{g}} \cdot \left(\mathrm{d}W_\tau^Q + \frac{\bar{u}}{\bar{g}}\,\mathrm{d}\tau\right) - \frac{1}{2}\int_0^T \frac{\|\bar{u}\|^2}{\bar{g}^2}\,\mathrm{d}\tau\right] \\ &= \mathbb{E}_Q\left[\int_0^T \frac{\bar{u}}{\bar{g}} \cdot \mathrm{d}W_\tau^Q\right] + \frac{1}{2}\,\mathbb{E}_Q\left[\int_0^T \frac{\|\bar{u}\|^2}{\bar{g}^2}\,\mathrm{d}\tau\right]. \end{aligned} \tag{23}$$

Since W_τ^Q is a Wiener process under Q, the stochastic integral $\int_0^T \frac{\bar{u}}{\bar{g}} \cdot \mathrm{d}W_\tau^Q$ is a martingale under Q with zero expectation. Substituting back the original variables via $\|\bar{u}(x, \tau)\|^2 = \|u(x, T - \tau)\|^2$ and $\bar{g}(\tau)^2 = g(T - \tau)^2$,

and changing the integration variable back to $t = T - \tau$, we obtain

$$D_{\mathrm{KL}}(p_{\mathrm{safe}}(0) \,\|\, p(0)) = \frac{1}{2}\,\mathbb{E}_{p_{\mathrm{safe}}(0)}\left[\int_0^T \frac{\|u(x,t)\|^2}{g(t)^2}\,\mathrm{d}t\right]. \tag{24}$$

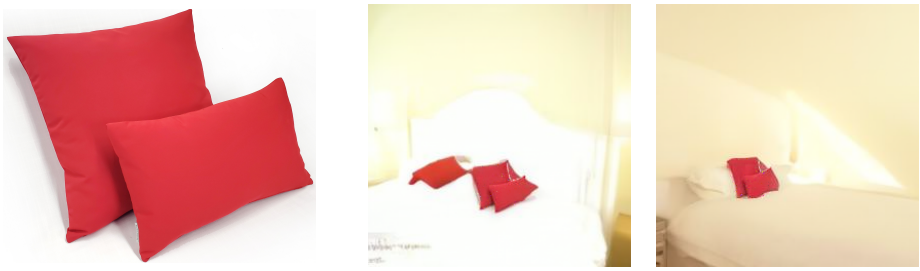
This concludes the proof. □

B Appendix: Additional experiments

B.1 Spatial content constraints with alternate reference images

Figure 6 demonstrates spatial and content-constrained image generation using red decorative pillows as the reference patch. Both constrained generations (b-c) preserve the pillows exactly at the specified location.

Notably, the generated scenes exhibit predominantly light tones. This occurs because the reference patch (a) includes white pixels near its boundaries, which the diffusion model interprets as local context. The spatially-varying mask $v(\mathbf{p})$ constrains boundary pixels to remain close to reference values, causing the model to generate ambient colors that transition naturally from the white boundary. This demonstrates that our CBF shield enforces constraints precisely as specified, including boundary characteristics. In practice, reference patches with neutral boundaries allow greater freedom in surrounding generation.



(a) (b) (c)

Figure 6: Spatial and content constrained generation with red pillow reference. (a) Reference patch (50×50 pixels) containing two red pillows with white background. (b) Generated image with 50 sampling steps. (c) Generated image with 200 sampling steps. Both preserve the pillows exactly while generating bedroom context. The white boundary pixels in the reference influence the surrounding color palette, resulting in light-toned ambient environments.

B.2 Constrained image generation with latent diffusion models

We investigate extending our CBF framework to latent diffusion models (Rombach et al., 2022), where the diffusion process operates in a compressed latent space rather than directly in pixel space. This extension is motivated by the scalability requirements of modern text-to-image models such as Stable Diffusion, which perform sampling in a low-dimensional latent space $\mathbb{R}^{4\times64\times64}$ that is decoded to images in $\mathbb{R}^{3\times512\times512}$ via a variational autoencoder (VAE).

Let $E : \mathbb{R}^n \to \mathbb{R}^d$ and $D : \mathbb{R}^d \to \mathbb{R}^n$ denote the VAE encoder and decoder, respectively, where $d \ll n$. Given a constraint defined in image space via a barrier function $h : \mathbb{R}^n \to \mathbb{R}$, we define the *latent barrier function* as the composition

$$\bar{h}(z) := h(D(z)), \tag{25}$$

which lifts the image-space constraint into the latent space via the decoder. The constricting barrier becomes $\tilde{h}(z,t) = \bar{h}(z) + \epsilon(z(T),t)$, and the CBF condition for the latent dynamics $\mathrm{d}z = [f_\theta(z,t) + u]\mathrm{d}t + g(t)\mathrm{d}w$ requires the gradient

$$\nabla_z \bar{h}(z) = \left(\frac{\partial D}{\partial z}\right)^\top \nabla_x h(x)\Big|_{x=D(z)}, \tag{26}$$

which is computed via a single backward pass through the decoder using automatic differentiation. The QP structure of Algorithm 1 is preserved, with the state space replaced by $\mathbb{R}^d$. We apply the spatial localization constraint from Section 5.2.1 to Stable Diffusion v1.5 (`sd-legacy/stable-diffusion-v1-5`) from Hugging Face Diffusers. The prompt to the model is "high quality image of a rustic bedroom". We constrain a rectangular region in the generated image to match a reference window patch, using the same per-pixel barrier function (14) evaluated on the decoded image $\hat{x} = D(z)$. The latent gradient (26) is backpropagated through the VAE decoder at each sampling step.

Figure 7 presents the results. The CBF-guided latent diffusion model responds to the spatial constraint: a visually distinct region appears at the specified location, indicating that the latent-space control input successfully biases the sampling trajectory toward the constrained region. However, the reference window patch is not preserved with pixel-level fidelity. The constraint region appears blurred, color-shifted, and blended into the surrounding scene, in contrast to the exact preservation achieved in pixel-space experiments (Figure 3 and 6). This degradation stems from a fundamental limitation that the VAE decoder D is not a diffeomorphism. The mapping from latent to image space is locally non-invertible and many-to-one, meaning that multiple latent codes can decode to similar but non-identical images. Consequently, enforcing $\bar{h}(z) = h(D(z)) \geq 0$ in latent space does not guarantee that the decoded image $\hat{x} = D(z)$ satisfies $h(\hat{x}) \geq 0$ with the same precision. The decoder introduces reconstruction artifacts that distort the constraint boundary, causing the gradient $\nabla_z \bar{h}$ to point in directions that are approximately but not exactly aligned with the true image-space constraint.

(a) (b)

Figure 7: Constrained image generation with latent diffusion models with prompt "high-quality image of a rustic bedroom". The constraint is to produce the reference window in Figure 3(a) in the top right corner, same as in Section 5.2.1. We see that the model responds to the constraint by producing a visually distinct region at the target location, but the reference patch is not preserved with pixel-level fidelity due to the VAE decoder's non-invertibility.

Formally, the safety guarantee of Theorem 4.1 holds with respect to the composed barrier $\bar{h}(z) = h(D(z))$: if $\tilde{h}(z(0), 0) \geq 0$, then $h(D(z(0))) \geq 0$. The issue is that the safe set $\bar{\mathcal{C}} = \{z \mid h(D(z)) \geq 0\}$ in latent space does not correspond exactly to the intended safe set in image space when the decoder has non-trivial reconstruction error. This is not a failure of the CBF framework itself, but rather a limitation of the representation space in which the diffusion process operates.