Trajectory Generation, Control, and Safety with Denoising Diffusion Probabilistic Models

Nicolò Botteghi [1] **Federico Califano** [2] **Mannes Poel** [3] **Christoph Brune** [1]

Abstract

We present a framework for safety-critical optimal control of physical systems based on denoising diffusion probabilistic models (DDPMs). The technology of control barrier functions (CBFs), encoding desired safety constraints, is used in combination with DDPMs to plan actions by iteratively denoising trajectories through a CBF-based guided sampling procedure. At the same time, the generated trajectories are also guided to maximize a future cumulative reward representing a specific task to be optimally executed. The proposed scheme can be seen as an offline and model-based reinforcement learning algorithm resembling in its functionalities a model-predictive control optimization scheme with receding horizon in which the selected actions lead to optimal and safe trajectories.

1. Introduction

When controlling a physical system, the concept of safety is crucial. In general, safety is not a system theoretic property (like e.g., stability) which possesses a precise mathematical definition, and safety requirements depend on the specific safety hazards that are considered. Systems for which some instance of safety is considered of paramount importance, are referred to as safety-critical systems. Examples include autonomous driving with adaptive cruise control and human-collision avoidance for collaborative robots (Ames et al., 2019). Safety constraints are conceptually distinct from the task-based specifications. For example, a collaborative robot performing a trajectory tracking task might be controlled using Lyapunov-based designs or learned control policies, but the underlying safety requirement of avoiding collisions with humans is transversal to any chosen controller, and must be always satisfied.

Control barrier functions (CBFs) (Ames et al., 2017; 2019) represent a formal framework aiming to achieve safety as a hard constraint in an optimization problem in which the cost function encodes information on the nominal task to be executed. In particular CBF-based safety constraints are represented by forward invariance of so-called safe sets, i.e. subsets of the state space which the controlled system should not leave during the task execution. We stress that within this context, safety becomes a mathematically rigorous system theoretic property and, even if unable to represent any possible safety hazard, it is very useful to design safety constraints, e.g. kinematic constraints for mechanical systems (Singletary et al., 2021).

Reinforcement learning (RL) (Sutton & Barto, 2018) has achieved outstanding success in control of dynamical and physical systems directly from high-dimensional sensory data (Kaelbling et al., 1996; Kaiser et al., 2019; Li, 2017; Botteghi et al., 2022). RL algorithms can be classified as either model-free or model-based. Model-free RL algorithms learn optimal policies directly from observations and rewards received by the agent from the environment after each action taken. On the other side, model-based algorithms first learn the environment dynamics, i.e. forward and reward models, and then use the learned models to generate samples for optimizing the policy. Independently of the chosen method, RL lacks of safety guarantees, which is one of the reasons why learning schemes are still poorly used in real safety-critical applications. As a consequence, embedding safety guarantees, for example using CBFs, in learning schemes is becoming a major research topic in the learning community, see (Dawson et al., 2023; Cheng et al., 2019; Ma et al., 2022) and references therein.

Intrinsically related to safety and CBFs, we find the need for a forward dynamical model able to accurately predict the evolution over time of the system given current state and control inputs. The forward model can be either derived from physical governing laws, especially for simple dynamical systems, or learned from data when limited or no physical knowledge is available (Brunton & Kutz, 2022).

[1]Mathematics of Imaging and AI, University of Twente, Enschede, The Netherlands [2]Robotics and Mechatronics, University of Twente, Enschede, The Netherlands [3]Datamanagement and Biometrics, University of Twente, Enschede, The Netherlands. Correspondence to: Nicolò Botteghi <n.botteghi@utwente.nl>.

Workshop on New Frontiers in Learning, Control, and Dynamical Systems at the International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA, 2023.

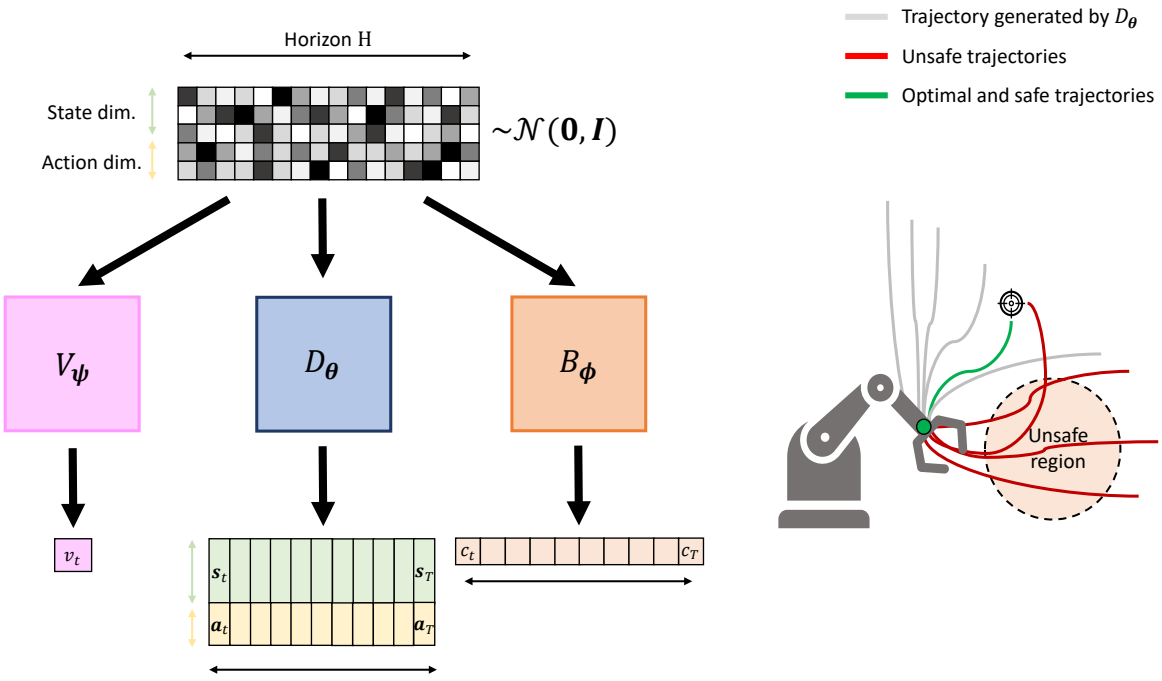


Figure 1. Proposed framework for planning and control using denoising diffusion probabilistic models. Generation from random noise of dynamically consistent state-action trajectories with horizon H from a generic timestep t to $T = t + H$.

The latter is often the case for complex dynamical systems (Brunton & Kutz, 2022) and the progress of deep learning (Lecun et al., 2015) has made learning, and controlling, dynamical systems from data increasingly popular in recent years. In most cases, data-driven methods focus on learning, for example via neural networks or Gaussian processes (Rasmussen, 2004), a (Markovian) 1-step-ahead forward model predicting the state at the next timestep given (history of) state(s) and action(s). However, when making long-term prediction, the 1-step-ahead forward model has to be iteratively called at the price of accumulating more and more prediction error the longer the horizon. An alternative solution was proposed by (Janner et al., 2022), where instead of learning a model predicting the next state, they introduce a trajectory-based generative model outputting sequences of state-action pairs, i.e. trajectories. It is worth mentioning that this model is only locally Markovian since each step of the trajectory is a function of the past and future state-action pairs. Therefore, the model has to be able to learn the causal relations from data, i.e. observed state-action pairs over time, to properly generate trajectories.

With reference to Figure 1, we frame our contribution in the context of safety-critical optimal control in an offline and model-based RL framework, proposing a scheme utilizing denoising diffusion probabilistic models (DDPMs) (Sohl-Dickstein et al., 2015; Ho et al., 2020) for optimal and safe trajectory generation. As hallmark of our contribution, three properties are approximated by three different DDPMs, and as such represent conceptually distinct units which can be independently changed in the proposed architecture:

- *dynamically consistent trajectory generation*, i.e. a diffusion model $D_{\boldsymbol{\theta}}$ is trained to generate trajectories consistent with the dynamics of the system,
- *value estimation*, i.e. a diffusion model V_{ψ} is trained to learn the expected return of each trajectory, encoding the specific task to be optimally executed, and
- *safety classification*, i.e., a diffusion model B_{ϕ} is trained to classify the generated trajectory as safe or unsafe, based on the CBF constraint.

Given these three DDPMs, the problem of safety-critical optimal control can be rephrased as a problem of conditional sampling of trajectories, i.e. sequence of state-action pairs, from $D_{\boldsymbol{\theta}}$ using V_{ψ} and B_{ϕ} as guides. With this method, the generated trajectories are not only optimal, but also safe.

Our approach closely relates to the work in (Janner et al., 2022), where two diffusion models, predicting trajectories and values, were proposed together with a conditional sampling scheme. However, differently from them we focus on safety-critical control with the addition of a third diffusion

model and a new conditional sampling scheme combining optimality and safety. After applying the first action of the trajectory, a new conditional sampling step takes place given the next state of the environment as initial state, resembling an optimization scheme with receding horizon. As a consequence, similarly to what has been achieved in (Zeng et al., 2021) in a model predictive control (MPC) framework, the CBF-based safety constraint influences the choice of the trajectory over the whole planning horizon, and not only at the current state, making the planning appoach not over-conservative. The view on the whole trajectory is the main conceptual differences with respect to (Cheng et al., 2019), also combining RL and CBFs, but where the CBF-based constraint is solved at every iteration to modify online the model-free RL policy.

2. Preliminaries

2.1. Control Barrier Functions

Control barrier functions (CBFs) are used to guarantee invariance of a set $\mathcal{C}$, normally referred to as *safe set*. The latter is designed to encode constraints depending on the state of the controlled system, e.g., obstacle avoidance for robot manipulators. We present the relevant background directly in its discrete-time formulation (Ahmadi et al., 2019; Zeng et al., 2021; Xiong et al., 2023) to conveniently integrate the technique in the developed learning scheme. Furthermore, we use the standard RL notation to indicate state, action, rewards, and value function (Sutton & Barto, 2018).

Consider a discrete-time system

$$\mathbf{s}_{t+1} = f(\mathbf{s}_t, \mathbf{a}_t) \tag{1}$$

where $\mathbf{s}_t \in \mathcal{S} \subset \mathbb{R}^n$ is the state at time step $t \in \mathbb{N}$, the control action $\mathbf{a}_t \in \mathcal{A}$ is applied to the system, and $f : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is a continuous map. The safe set $\mathcal{C}$ is defined as

$$\begin{aligned} \mathcal{C} &= \{\mathbf{s} \in \mathcal{S} : h(\mathbf{s}) \geq 0\}, \\ \partial\mathcal{C} &= \{\mathbf{s} \in \mathcal{S} : h(\mathbf{s}) = 0\}, \\ \text{Int}(\mathcal{C}) &= \{\mathbf{s} \in \mathcal{S} : h(\mathbf{s}) > 0\}, \end{aligned}$$

where $h : \mathcal{S} \to \mathbb{R}$ is a continuous map. Remind that a class $\mathcal{K}$ function is a strictly increasing function $\alpha : [0, a) \to [0, \infty)$ with $\alpha(0) = 0$ and $a > 0$. The function $h(\mathbf{s})$ is then defined to be a CBF if there exists a class $\mathcal{K}$ function α satisfying $\alpha(u) < u, \forall u > 0$ such that

$$h(\mathbf{s}_{t+1}) - h(\mathbf{s}_t) \geq -\alpha(h(\mathbf{s}_t)), \quad \forall \mathbf{s} \in \mathcal{S}. \tag{2}$$

The set $\mathcal{C}$ is called forward invariant if $\mathbf{s}_0 \in \mathcal{C}$ implies $\mathbf{s}_t \in \mathcal{C}, \forall t \in \mathbb{N}$. The following theorem, which represents the core of CBFs in the current context, relates existence of CBF and forward invariance of $\mathcal{C}$ achieved by means of control.

Theorem 2.1. *(Ahmadi et al., 2019) Consider system (1) and the previously defined set $\mathcal{C}$. Any control policy $\pi(\mathbf{a}_t|\mathbf{s}_t)$ implementing $\mathbf{a}_t$ in a way that (2) is satisfied will render $\mathcal{C}$ forward invariant.*

Remark 2.2. We stress that in the literature different types of CBFs are present. These vary, both in continuous and discrete time, on the basis of i) the specific set $\mathcal{S}$ where condition (2) must hold, which can be a proper subset of the total state space (in this case the CBFs are normally referred to explicitly as "CBFs on $\mathcal{S}$"); and ii) whether the CBF diverges or goes to 0 at the boundary $\partial\mathcal{C}$ of the safe set. In the latter case, which is the one treated in this work, CBFs are sometimes referred to as "zeroing" CBFs, and have the advantage to be well defined outside the safe set. It follows (see proof Theorem 1 in (Ahmadi et al., 2019)) that existence of zeroing CBFs on $\mathcal{S}$ not only achieve forward invariance, but also asymptotic stability of $\mathcal{C}$.

Remark 2.3. Condition (2) for an autonomous discrete time system $\mathbf{s}_{t+1} = f(\mathbf{s}_t)$ is actually necessary and sufficient (i.e., equivalent) to achieve forward invariance of the set $\mathcal{C}$ (see Theorem 1 in (Ahmadi et al., 2019)). It follows that finding a control policy applied to (1) which satisfies (2) is also necessary and sufficient (i.e., equivalent) to produce forward invariance of $\mathcal{C}$ in the closed-loop system. An interesting consequence is that the discrete time CBF approach is not conservative (but in fact equivalent) to achieve forward invariance of a set.

Remark 2.4. Normally one chooses the linear class $\mathcal{K}$ function $\alpha(u) = \lambda u$, with $\lambda \in \mathbb{R},\ 0 \leq \lambda \leq 1$. In this way it holds $h(\mathbf{s}_t) \geq (1-\lambda)^t h(\mathbf{s}_0)$, which clearly shows forward invariance of $\mathcal{C}$ for every admissible λ. The constraint (2) with the choice $\lambda = 1$ corresponds to achive invariance in the less conservative case, i.e., $h(\mathbf{s}_t) \geq 0, \forall t \in \mathbb{N}$, while the choice $\lambda = 0$ collapses (2) into a Lyapunov condition achieving $h(\mathbf{s}_t) \geq h(\mathbf{s}_{t-1}), \forall t \in \mathbb{N}$.

The way discrete-time CBFs are implemented in practice is by casting condition (2) as a constraint of a discrete optimisation problem, which in its basic forms can be represented as

$$\begin{aligned} \mathbf{a}_t^* = \text{argmin}_{\mathbf{a}_t \in \mathcal{A}} \quad & \mathbf{J}(\mathbf{s}_t, \mathbf{a}_t) \\ \text{s.t.} \quad & h(f(\mathbf{s}_t, \mathbf{a}_t)) - h(\mathbf{s}_t) \geq -\alpha(h(\mathbf{s}_t)) \end{aligned} \tag{3}$$

The level of sophistication of the latter can vary on the basis of the application, by adding other constraints (e.g., input constraints), and/or slack variables to help the feasibility of the problem.

2.2. Denoising Diffusion Probabilistic Models

Denoising diffusion probabilistic models (DDPMs) (Sohl-Dickstein et al., 2015; Ho et al., 2020) are a class of (deep) probabilistic generative models, such as generative adversarial networks (Goodfellow et al., 2020), variational autoencoders (Kingma & Welling, 2014), and score-matching models (Song & Ermon, 2019; 2020) learning generative data distributions out of which we can sample new data. DDPMs are parametrized Markov chains generating data from random noise by performing a step-wise denoising of the random vectors. The denoising process, often referred to as the reverse process, is learned from data with the goal of reverting the forward process of gradually adding noise to the data. The forward process can be written as:

$$\begin{aligned} q(\mathbf{x}^{1:K}|\mathbf{x}^0) &= \prod_{k=1}^{K} q(\mathbf{x}^k|\mathbf{x}^{k-1}), \\ q(\mathbf{x}^k|\mathbf{x}^{k-1}) &= \mathcal{N}(\sqrt{1-\beta_k}\mathbf{x}^{k-1}, \beta_k\mathbf{I}) \end{aligned} \tag{4}$$

where $\mathbf{x} \sim \mathcal{D}$ is a data sample from a dataset $\mathcal{D}$, the superscript $k = 1, \dots, K$ indicates the k^{th} diffusion step, K the number of diffusion steps, $\mathbf{x}^0$ the noise-free data, and $\beta_k \in (0, 1)$ is an hyperparameter controlling the variance of the forward diffusion process. The reverse process is described by:

$$\begin{aligned} p_{\boldsymbol{\theta}}(\mathbf{x}^{0:K}) &= \prod_{k=1}^{K} p_{\boldsymbol{\theta}}(\mathbf{x}^{k-1}|\mathbf{x}^k) \\ p_{\boldsymbol{\theta}}(\mathbf{x}^{k-1}|\mathbf{x}^k) &= \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}^k, k), \boldsymbol{\Sigma}^k) \end{aligned} \tag{5}$$

where $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}^k, k)$ is the mean learned by the diffusion model, $\boldsymbol{\Sigma}^k$ is the covariance matrix usually following a cosine schedule (Nichol & Dhariwal, 2021) (and not learned from data), and the subscript $\boldsymbol{\theta}$ indicates the (learnable) parameters of the diffusion model.

DDPMs can be trained by maximizing the variational lower bound similarly to variational autoencoders (Sohl-Dickstein et al., 2015), or using the simplified objective introduced by (Ho et al., 2020):

$$\begin{aligned} L(\boldsymbol{\theta}) &= \mathbb{E}_{k,\epsilon,\mathbf{x}_0}[||\epsilon^k - \epsilon_{\boldsymbol{\theta}}(\mathbf{x}^k, k)||^2] \\ &= \mathbb{E}_{k,\epsilon,\mathbf{x}_0}[||\epsilon^k - \epsilon_{\boldsymbol{\theta}}(\sqrt{\bar{\alpha}_k}\mathbf{x}^0 + \sqrt{1-\bar{\alpha}_k}\epsilon^k, k)||^2] \end{aligned} \tag{6}$$

where $k \sim \mathcal{U}\{1, \dots, K\}$ with $\mathcal{U}$ indicating a uniform distribution, $\epsilon^k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the target noise, $\bar{\alpha}_k = \prod_{k=1}^{K} \alpha_k$ with $\alpha_k = 1 - \beta_k$, and with the learnable mean rewritten as:

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}^k, k) = \frac{1}{\sqrt{\alpha_k}}(\mathbf{x}^k - \frac{1-\alpha_k}{\sqrt{1-\bar{\alpha}_k}}\epsilon_{\boldsymbol{\theta}}(\mathbf{x}^k, k)) \tag{7}$$

To further improve and control the generation process, it is possible to introduce a classifier $p_{\boldsymbol{\phi}}(y|\mathbf{x}^k)$ and use its gradient $\nabla_{\mathbf{x}^k} p_{\boldsymbol{\phi}}(y|\mathbf{x}^k)$ with respect to the input $\mathbf{x}^k$ to guide the reverse diffusion process towards samples from a class y. This procedure is commonly referred to as classifier-guided sampling (Dhariwal & Nichol, 2021).

2.2.1. Planning and Control

Instead of considering a generic data $\mathbf{x}$, we assume we have recorded trajectories of our dynamical system $\boldsymbol{\tau}_t = (\mathbf{s}_t, \mathbf{a}_t, \dots, \mathbf{s}_T, \mathbf{a}_T)$, where $\mathbf{s}_t$ corresponds to the state vector at timestep t, $\mathbf{a}_t$ to the control action, and the subscript $t \in \{0, ..., T\}$ to the actual timestep of agent-environment interaction of the underlying RL scheme, while the superscript k indicates the diffusion steps as introduced in Section 2.2. Given a sampled trajectory $\boldsymbol{\tau}^K \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, it is possible to use a trained diffusion model to reverse the forward process and to generate a trajectory $\boldsymbol{\tau}_t$ consistent with the observed dynamics of the system, as shown in (Janner et al., 2022):

$$p_{\boldsymbol{\theta}}(\boldsymbol{\tau}_t^{k-1}|\boldsymbol{\tau}_t^k) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{\tau}_t^k, k), \boldsymbol{\Sigma}^k) \tag{8}$$

In a classic model-based RL fashion, one may sample trajectories from the reverse diffusion model, apply the control actions to the real environment and evaluate the value of each visited state. However, analogously to the classifier-guided sampling introduced by (Dhariwal & Nichol, 2021), diffusion models can be used to perform conditional sampling, e.g., sampling images from a specific class, by using a classifier to adjust the mean of the reverse process during the diffusion steps. The use of conditional sampling for control was first explored by (Janner et al., 2022) by transforming the problem of optimal trajectory generation in a problem of conditional sampling:

$$\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{\tau}) \approx p_{\boldsymbol{\theta}}(\boldsymbol{\tau}) f(\boldsymbol{\tau}) \tag{9}$$

where $f(\boldsymbol{\tau}_t)$ may encode prior information, desired goals, or general reward and cost functions to optimize. In particular, we can use a trained value function model, indicated by $V_{\boldsymbol{\psi}}$ to sample trajectories maximizing the value function, i.e. optimal trajectories. A single guided-diffusion step can be written as:

$$\boldsymbol{\tau}^{k-1} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}^{k-1} + \eta_1 \boldsymbol{\Sigma}^{k-1} \nabla_{\boldsymbol{\tau}^k} V_{\boldsymbol{\psi}}(\boldsymbol{\mu}_{\boldsymbol{\theta}}), \boldsymbol{\Sigma}^{k-1}) \tag{10}$$

with η_1 a scaling constant, and $\boldsymbol{\mu}_{\boldsymbol{\theta}} = \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{\tau}_t^k, k)$ for the sake of lightening the notation. After the generation the planning, the first action is executed in the environment, and similarly to MPC methods with receding horizon, the plan is generated again given the next state of the environment as initial state of the trajectory.

While this planning strategies allows the generation of optimal trajectories, differently from (Janner et al., 2022), in our work we focus on the problem of optimal planning under safety constraint where we want to generate not only optimal, but also safe trajectories.

3. Optimal and Safe Trajectory Generation with Denoising Diffusion Probabilistic Models

In this work, we study the problem optimal and safe trajectory generation using DDPMs and conditional sampling. With reference to Figure 1, we decompose the problem of optimal and safe trajectory planning into the problem of learning:

- $D_{\boldsymbol{\theta}}$ generating trajectories $\boldsymbol{\tau}_t = (\mathbf{s}_t, \mathbf{a}_t, \dots, \mathbf{s}_T, \mathbf{a}_T)$ from random noise consistent with the measurement data of the dynamical system considered,
- $V_{\boldsymbol{\psi}}$ estimating the value, i.e. expected discounted return, of each trajectory, again from random noise, and
- $B_{\boldsymbol{\phi}}$ classifying whether each timestep of the trajectories generated is safe or unsafe in accordance with the CBF-based safety constraint, defined by Equation (2).

These three models are DDPMs (see Section 2.2) with architecture described in details in Appendix A.

While it is possible to train a single DDPM to generate trajectories, predicting values, and assessing safety of the state-action pairs, the independence of the three models is important for data efficiency and flexibility of use. For example, if a robot is moved to a new room the safety regions may change, while its dynamics or task may not. With our framework, we just need to retrain the safety classifier $B_{\boldsymbol{\phi}}$. The same holds if we change reward function, with the difference that we need now to retrain only the value function model $V_{\boldsymbol{\psi}}$.

3.1. Planning as Conditional Sampling

In this section, we describe the conditional sampling procedure for the generation of optimal and safe trajectories. Given $D_{\boldsymbol{\theta}}$, we can generate trajectories from random noise accordingly to Equation (8). The generation of the trajectories can be conditioned on initial and/or final states of the trajectories through the process called inpainting (Sohl-Dickstein et al., 2015), i.e. the problem of consistently filling the gaps or a trajectory given some known states or actions (see Figure 2 for an example). Since we focus on path planning with unknown target location, in our experiments we condition the generated plan only on the initial state. Using $V_{\boldsymbol{\psi}}$, we can guide $D_{\boldsymbol{\theta}}$ toward the generation of high-value trajectories according to Equation (10), i.e. trajectories reaching the target (Janner et al., 2022), yet possibly unsafe. The guided planning procedure is described in Algorithm 1, where the guide function is generically indicated by f.

To make these high-value trajectories safe, accordingly to the CBFs methodology, one may solve the constrained optimization problem in Equation (3) for each step of the trajectory in order to "filter" the actions. However, the problem in (3) is expensive to solve, especially for non-affine dynamics and/or constraints (Cheng et al., 2019), and has to be computed every time before applying an action to the environment.

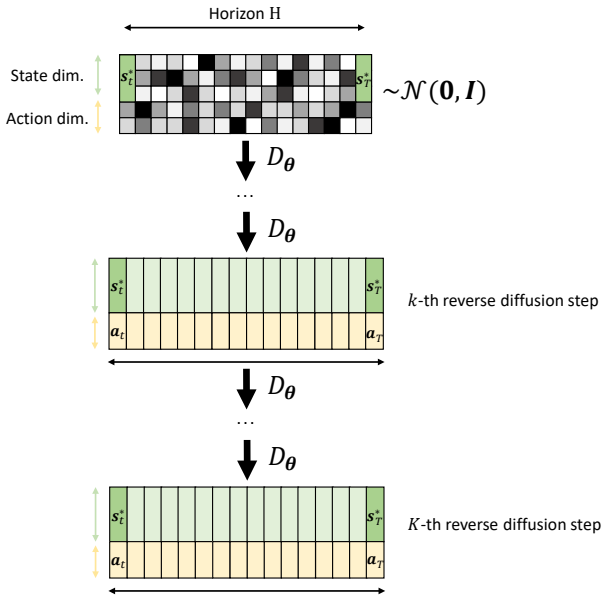


Figure 2. Reverse diffusion process with inpainting of initial and final states of the trajectory indicated with $\mathbf{s}_t^*$ and $\mathbf{s}_T^*$.

Here, we propose an alternative solution. Recall that $B_{\boldsymbol{\phi}}$ is able to classify whether each step of a trajectory is safe or not. Similarly to the classifier-guide sampling method (Dhariwal & Nichol, 2021), often employed in the process of generating high-quality images, we can guide $D_{\boldsymbol{\theta}}$ to generate only trajectories belonging to the "safe" class by adjusting the mean of the reverse process accordingly to:

$$\boldsymbol{\tau}^{k-1} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}^{k-1} + \eta_2 \boldsymbol{\Sigma}^{k-1} \nabla_{\boldsymbol{\tau}^k} \log p_{\boldsymbol{\phi}}(\mathbf{c}|\boldsymbol{\tau}^k), \boldsymbol{\Sigma}^{k-1}) \tag{11}$$

where η_2 is a scaling constant, $p_{\boldsymbol{\phi}}(\mathbf{c}|\boldsymbol{\tau}^k) = B_{\boldsymbol{\phi}}$, and $\mathbf{c} = (c_t, \cdots, c_T)$ is the vector containing the class labels for each step of the trajectory. We can use the sampling procedure in Equation (11) to generate trajectories the are classified as safe by $B_{\boldsymbol{\phi}}$. Again if both initial and final positions are known, one can simply use (11) together with inpainting to generate safe trajectories connecting the two points. However, these trajectories may not be optimal in terms of reward accumulated, e.g. may be over-conservative or require high-control effort.

Instead of guiding the planning towards optimality or safety, we propose to do both by employing both guides at the same

Algorithm 1 Guided Planning

Input: $D_{\boldsymbol{\theta}}$, and guide f
for $i = 1$ **to** I **do**
  Reset environment and sample $\mathbf{s}_1$
  for $j = 1$ **to** J **do**
    Sample plan $\boldsymbol{\tau}^K \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ conditioned on $\mathbf{s}_j$
    for $k = K$ **to** 1 **do**
      $\boldsymbol{\mu}_{\boldsymbol{\theta}} \leftarrow D_{\boldsymbol{\theta}}(\boldsymbol{\tau}^k)$
      $\boldsymbol{\tau}^{k-1} \leftarrow\sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}} + \eta\boldsymbol{\Sigma}\nabla f, \boldsymbol{\Sigma}^k)$
      $\boldsymbol{\tau}^{k-1}_{\mathbf{s}_j} \leftarrow \mathbf{s}_j$
    end for
    Apply first action of the plan to the environment
    Observe next state $\mathbf{s}_{j+1}$, reward r_j, terminal condition d_{j+1}, and safety constraint c_j
    $\mathbf{s}_j \leftarrow \mathbf{s}_{j+1}$
  end for
end for

time and introducing a new conditional sampling procedure:

$$\boldsymbol{\tau}^{k-1} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}(\eta_1 \nabla f_1 + \eta_2 \nabla f_2), \boldsymbol{\Sigma}) \tag{12}$$

where $f_1 = V_{\boldsymbol{\psi}}(\boldsymbol{\tau}^k)$ and $f_2 = \log p_{\boldsymbol{\phi}}(\mathbf{c}|\boldsymbol{\tau}^k)$[1]. Similarly to before, we can now generate optimal and safe trajectories by simply plugging (12) in Algorithm 1.

DDPMs are probabilistic models, thus the generated trajectories may differ every time we call Algorithm 1. Therefore, we do not simply generate a single trajectory but a batch of them (64 in our experiments), we then select one accordingly to a predefined criterium, and we apply the first action of the plan to the environment. Because our main concern is safety, we select the trajectory without collisions, but more advance criteria may be used, such as considering a weighted sum of value and collisions or even learning the weighting factor. However, we leave this interesting research direction to future work.

3.2. Training Objectives

To train our models and test their generalization capabilities in low-data regimes, we collect a dataset $\mathcal{D}$ with 300 trajectories of 100 steps each for a total of 30000 tuples $(\mathbf{s}_t^{\text{target}}, \mathbf{a}_t^{\text{target}}, \mathbf{s}_{t+1}^{\text{target}}, r_t^{\text{target}}, d_t^{\text{target}}, c_t^{\text{target}})$, where r_t^{target} is the instantaneous reward, d_t^{target} is a boolean flag indicating the end of the episode, and:

$$c_t^{\text{target}} = \begin{cases} 0 & \text{if} \quad h(\mathbf{s}_{t+1}) - h(\mathbf{s}_t) \geq -\alpha(h(\mathbf{s}_t)) \\ 1 & \text{otherwise} \end{cases} \tag{13}$$

is the label generated through the use of the CBF constraint in Equation (2). We use a random control policy to generate the training dataset.

[1]We drop subscripts and superscripts for the sake of conciseness.

Similarly to (Janner et al., 2022), instead of relying of the objective function in Equation (6)[2], we train the diffusion model $D_{\boldsymbol{\theta}}$ using the mean-squared error loss between the generated trajectory $\boldsymbol{\tau}$ and the measured one $\boldsymbol{\tau}^{\text{target}}$:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\tau}_t, \boldsymbol{\tau}_t^{\text{target}} \sim \mathcal{D}}[||\boldsymbol{\tau}_t - \boldsymbol{\tau}_t^{\text{target}}||_2] \tag{14}$$

The value function model is instead trained to predict the expected return of a given trajectory, but again with mean-squared error loss:

$$\mathcal{L}(\boldsymbol{\psi}) = \mathbb{E}_{\boldsymbol{\tau}_t, r_{t:H} \sim \mathcal{D}}[||v_t - v_t^{\text{target}}||_2] \tag{15}$$

where $v_t^{\text{target}} = \sum_{t=0}^{H-1} \gamma^t r_t^{\text{target}}$ with H equal to the planning horizon and $\gamma \in [0, 1]$ is the discount factor. Eventually, the CBF classifier $B_{\boldsymbol{\phi}}$ is trained with the cross entropy loss:

$$\mathcal{L}(\boldsymbol{\phi}) = \mathbb{E}_{\boldsymbol{\tau}_t, c_{t:T} \sim \mathcal{D}}[\sum_{t=0}^{T-1} -c_t^{\text{target}} \log(c_t) + (1 - c_t^{\text{target}}) \log(c_t)] \tag{16}$$

4. Numerical Experiments

For our numerical experiments [3], we simulate a two degrees of freedom, fully-actuated, planar manipulator in the task of reaching a randomly-sampled target from a randomly-sampled initial configuration of its angles and angular velocities. Additionally, we add a circular unsafe region, the end-effector must not enter into as shown in Figure 3.

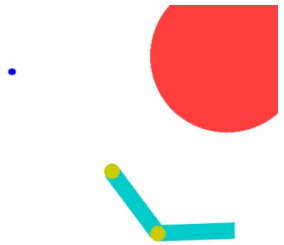


Figure 3. Openai Gym simulated environment. The blue dot represents the target location that the end-effector has to reach, while the red circle represent the unsafe region.

The state is composed of the joint angles α_1 and α_2, expressed with their sine and cosine, their angular velocity $\dot{\alpha}_1$ and $\dot{\alpha}_2$, and the target position (x_{tg}, y_{tg}) for a total dimension of 8. The actions a_1, a_2 are the torques at the two joints

[2]For the example considered, we did not notice substantial difference between training the models to reconstruct the original signal or to reconstruct the noise.

[3]The framework is implemented using PyTorch (Paszke et al., 2019) and Openai Gym (Brockman et al., 2016). The code can be found at: `github.com/nicob15`.

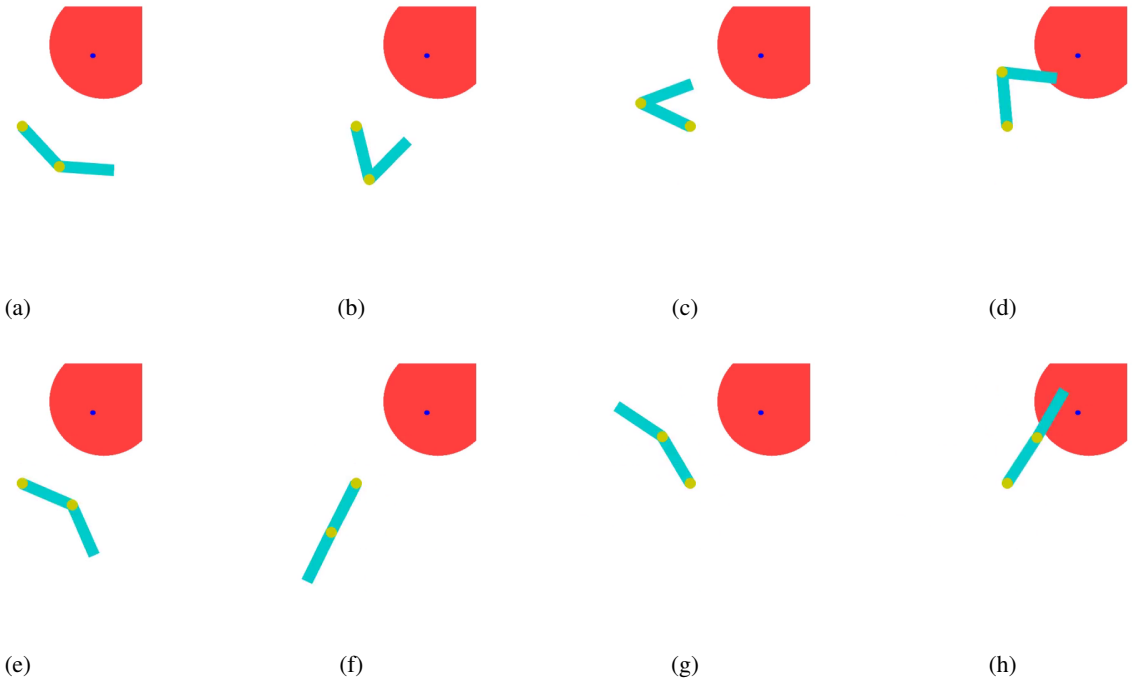


Figure 4. Trajectory generated by the diffusion planner with value-function guide.

$\in [-1.0, 1.0]$. Because the goal of the agent is to reach a target position starting from any initial configuration, we chose as the reward function the negative of the Euclidean distance from the end-effector to the target:

$$R(\mathbf{s}) = -d(x_{ee}, y_{ee}, x_{tg}, y_{tg}) \tag{17}$$

with x_{ee}, y_{ee} indicating the coordinates of the end-effector, and d indicating the Euclidean distance. This reward function, when maximized, corresponds to the end-effector reaching the target.

To assess the performance of this framework in low data regimes, we first train offline $D_{\boldsymbol{\theta}}$, $V_{\boldsymbol{\psi}}$, and $B_{\boldsymbol{\phi}}$ using a small dataset of randomly generated trajectory (as described in Section 3.2) with loss functions (14)-(16) respectively and then we evaluate the policy generated using the guided planning generated Algorithm 1 using only the value function guide on a set of 100 randomly-generated target. We record success rate[4], average rewards, average number of steps, and standard deviation of them. We compare with soft-actor critic (SAC) (Haarnoja et al., 2018) trained for 100K, 200K, and 300K iteration for 300 episodes of 100 steps each for the sake of a fair comparison. We report the results in Table 1. The complete list of hyperparameters and an ablation study on the guide scale η_1 can be find respectively in Appendix B and C.

Secondly, we want to test whether we can generate optimal and safe trajectories using Algorithm 1 by using both the value and the safety-classifier guide accordingly to Equation (12).

5. Results

From Table 1, we can notice that the proposed offline and model-based approach in low-data regimes is able to generalize better that model-free SAC in terms of success rate and average number of steps to reach the target, but not in term of average reward. The small difference is likely due to the different dataset used for training the two methods. Despite the two methods are trained on the same amount of data, the diffusion models are trained completely offline on randomly generated trajectories, while SAC is trained online. Therefore, SAC is able to experience more high-reward trajectories in later episodes.

Method	Succ. Rate	Avg. Rew.	Avg. Steps
$D_{\boldsymbol{\theta}} + V_{\boldsymbol{\psi}}$ (200K it.)	0.65	-104.24 ± 78.06	53.42 ± 38.42
SAC (300K it.)	0.57	-98.67 ± 74.44	55.97 ± 40.89
SAC (200K it.)	0.56	-102.3 ± 68.93	59.01 ± 38.09
SAC (100K it.)	0.55	-107.07 ± 69.18	62.71 ± 38.59

Table 1. Comparison of the proposed offline, model-based approach trained for 200K iteration and online, model-free SAC trained for 300K, 200K, 100K iterations on a set of 100 randomly-generated target and initial conditions.

[4]We consider success when the end-effect reaches the target, with a predefined tolerance ϵ, withing 100 steps, and unsuccess otherwise.

In Figure 4 and 5, we show the trajectories generated by the DDPMs with value guide (Equation (10)) and with the proposed value and safety-classifier guide (Equation (12)). It is worth mentioning that the dataset used to train the models does not contain targets in the unsafe region, as in this case. Therefore, this is a good example of generalization capabilities of the diffusion models even in low-data regimes. While the diffusion planner can reach the target (see Figure 4), without any safety constraint it naturally ends up in the unsafe region multiple times. On the other side, the diffuser planner employing the conditional sampling strategy in (12) does not violate the safery region while close to the (unreachable) target. Additional results can be found in Appendix D.

6. Conclusion

In this work, we studied the problem of safety-critical optimal control using data-driven generative models, i.e. DDPMs. In particular, we transform the problem of learning an optimal and safe policy into the problem of conditional sampling of trajectories, i.e. state-action pairs over an horizon H, either with high value and safe using a value function model and a safety (CBF-inspired) classifier as guides. We show that the method is not only more sample efficient that model-free RL in low-data regimes, but it can also be used to generate trajectories avoiding the unsafe regions. Additionally, we decoupled trajectory generation, value estimation, and safety prediction making by using three independent DDPMs, making the framework extremely flexible when dealing with changes in the environment, in the task, or in the safety regions.

References

Ahmadi, M., Singletary, A., Burdick, J. W., and Ames, A. D. Safe Policy Synthesis in Multi-Agent POMDPs via Discrete-Time Barrier Functions. *Proceedings of the IEEE Conference on Decision and Control*, 2019-Decem:4797–4803, 2019. ISSN 25762370. doi:10.1109/CDC40024.2019.9030241.

Ames, A. D., Xu, X., Grizzle, J. W., and Tabuada, P. Control Barrier Function Based Quadratic Programs for Safety Critical Systems. *IEEE Transactions on Automatic Control*, 62(8):3861–3876, 2017. ISSN 00189286. doi:10.1109/TAC.2016.2638961.

Ames, A. D., Coogan, S., Egerstedt, M., Notomista, G., Sreenath, K., and Tabuada, P. Control barrier functions: Theory and applications. *2019 18th European Control Conference, ECC 2019*, pp. 3420–3431, 2019. doi:10.23919/ECC.2019.8796030.

Botteghi, N., Poel, M., and Brune, C. Unsupervised Representation Learning in Deep Reinforcement Learning: A Review. 8 2022. URL http://arxiv.org/abs/2208.14226.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. OpenAI Gym. 2016. URL http://arxiv.org/abs/1606.01540.

Brunton, S. L. and Kutz, N. J. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. 2022. URL https://books.google.nl/books?hl=en&lr=&id=rxNkEAAAQBAJ&oi=fnd&pg=PR9&dq=Data+Driven+Science+and+Engineering&ots=kmG_U1Jx3p&sig=QbxQQ6OrTC3qAcRwEccAMgNqS4U&redir_esc=y#v=onepage&q=Data%20Driven%20Science%20and%20Engineering&f=false.

Cheng, R., Orosz, G., Murray, R. M., and Burdick, J. W. End-to-End Safe Reinforcement Learning through Barrier Functions for Safety-Critical Continuous Control Tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3387–3395, 7 2019. ISSN 2374-3468. doi:10.1609/AAAI.V33I01.33013387. URL https://ojs.aaai.org/index.php/AAAI/article/view/4213.

Dawson, C., Gao, S., and Fan, C. Safe Control With Learned Certificates: A Survey of Neural Lyapunov, Barrier, and Contraction Methods for Robotics and Control. *IEEE Transactions on Robotics*, pp. 1–19, 2023. ISSN 1552-3098. doi:10.1109/TRO.2022.3232542.

Dhariwal, P. and Nichol, A. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, volume 11, pp. 8780–8794, 2021. ISBN 9781713845393.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative Adversarial Networks. *COMMUNICATIONS OF THE ACM*, 63(11), 2020. doi:10.1145/3422622.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., and Levine, S. Soft Actor-Critic Algorithms and Applications. 2018. URL http://arxiv.org/abs/1812.05905.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 2020-Decem, 2020. URL https://github.com/hojonathanho/diffusion.

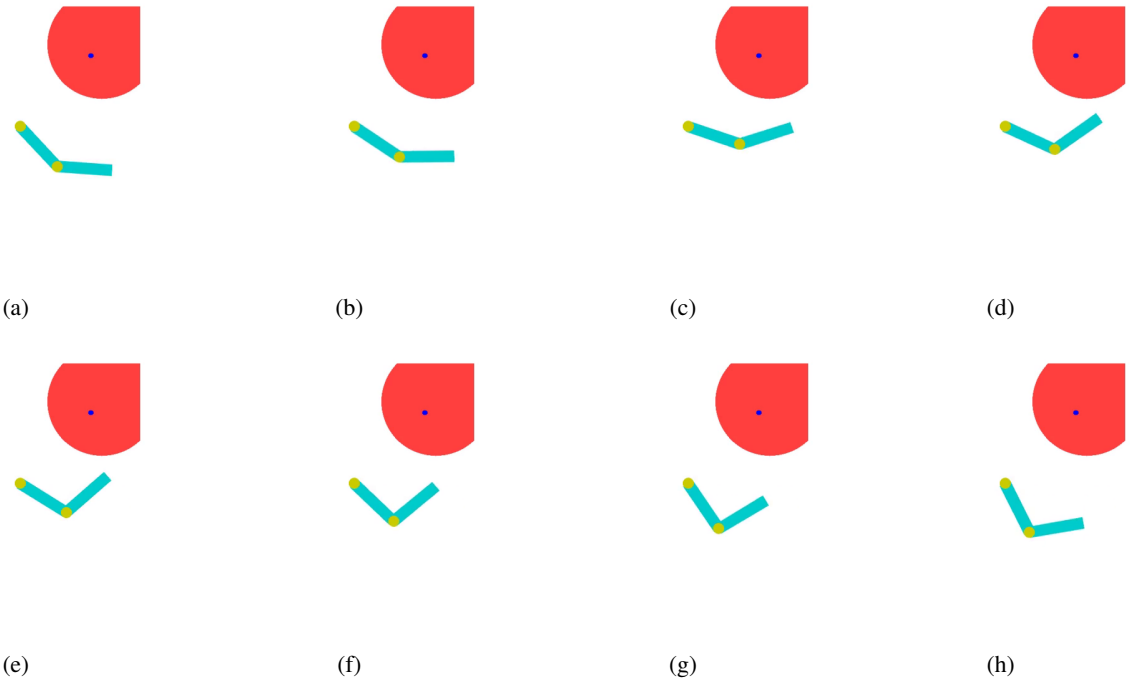


Figure 5. Trajectory generated by the diffusion planner with value-function and safety-classifier guides.

Janner, M., Du, Y., Tenenbaum, J. B., and Levine, S. Planning with Diffusion for Flexible Behavior Synthesis. 2022. URL http://arxiv.org/abs/2205.09991.

Kaelbling, L. P., Littman, M. L., and Moore, A. W. Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, 4:237–285, 5 1996. ISSN 1076-9757. doi:10.1613/JAIR.301. URL https://www.jair.org/index.php/jair/article/view/10166.

Kaiser, L., Babaeizadeh, M., Milos, P., Osinski, B., Campbell, R. H., Czechowski, K., Erhan, D., Finn, C., Kozakowski, P., Levine, S., Mohiuddin, A., Sepassi, R., Tucker, G., and Michalewski, H. Model-Based Reinforcement Learning for Atari. In *Reinforcement Learning for Cyber-Physical Systems*, pp. 69–92. 3 2019. doi:10.1201/9781351006620-4. URL http://arxiv.org/abs/1903.00374.

Kingma, D. P. and Ba, J. L. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 2014.

Lecun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature 2015 521:7553*, 521(7553):436–444, 5 2015. ISSN 1476-4687. doi:10.1038/nature14539. URL https://www.nature.com/articles/nature14539.

Li, Y. Deep Reinforcement Learning: An Overview. 1 2017. URL https://arxiv.org/abs/1701.07274v6.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019*, 2019. URL https://github.com/loshchil/AdamW-and-SGDW.

Ma, H., Liu, C., Li, S. E., Zheng, S., Chen, J., Firoozi, R., Mehr, N., Yel, E., Antonova, R., Bohg, J., Schwager, M., and Kochenderfer, M. Joint Synthesis of Safety Certificate and Safe Control Policy Using Constrained Reinforcement Learning, 5 2022. ISSN 2640-3498. URL https://proceedings.mlr.press/v168/ma22a.html.

Misra, D. Mish: A Self Regularized Non-Monotonic Activation Function. 8 2019. URL https://arxiv.org/abs/1908.08681v3http://arxiv.org/abs/1908.08681.

Nichol, A. and Dhariwal, P. Improved Denoising Diffusion Probabilistic Models. 2021. URL

https://github.com/openai/http://arxiv.org/abs/2102.09672.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury Google, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Xamla, A. K., Yang, E., Devito, Z., Raison Nabla, M., Tejani, A., Chilamkurthy, S., Ai, Q., Steiner, B., Facebook, L. F., Facebook, J. B., and Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. 2019.

Rasmussen, C. E. Gaussian Processes in Machine Learning. pp. 63–71. 2004. doi:10.1007/978-3-540-28650-9_4. URL http://link.springer.com/10.1007/978-3-540-28650-9_4.

Singletary, A., Kolathaya, S., and Ames, A. D. Safety-Critical Kinematic Control of Robotic Systems. *Proceedings of the American Control Conference*, 2021-May:14–19, 2021. ISSN 07431619. doi:10.23919/ACC50511.2021.9482954.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *32nd International Conference on Machine Learning, ICML 2015*, volume 3, pp. 2246–2255, 2015. ISBN 9781510810587.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems*, volume 2020-Decem, 2020.

Sutton, R. S. and Barto, A. G. Reinforcement Learning: An introduction, 2018. URL https://mitpress.ublish.com/ebook/reinforcement-learning-an-introduction-2-preview/2351/Cover.

Wu, Y. and He, K. Group Normalization. *International Journal of Computer Vision*, 128(3):742–755, 2020. ISSN 15731405. doi:10.1007/s11263-019-01198-w.

Xiong, Y., Zhai, D.-H., Tavakoli, M., and Xia, Y. Discrete-Time Control Barrier Function: High-Order Case and Adaptive Case. *IEEE Transactions on Cybernetics*, 53(5):3231–3239, 5 2023. ISSN 2168-2267. doi:10.1109/TCYB.2022.3170607.

Zeng, J., Zhang, B., and Sreenath, K. Safety-Critical Model Predictive Control with Discrete-Time Control Barrier Function. *Proceedings of the American Control Conference*, 2021-May:3882–3889, 2021. ISSN 07431619. doi:10.23919/ACC50511.2021.9483029.

A. Architecture

We use a similar architecture to (Janner et al., 2022) for the diffusion models $D_{\boldsymbol{\theta}}$, value function $V_{\boldsymbol{\psi}}$, safety classifier $B_{\boldsymbol{\phi}}$. The three models use a U-net with 6 residual blocks with linear attention. Each block is composed of two 1D temporal convolutions, group normalization (Wu & He, 2020) and Mish activation (Misra, 2019). The timesteps of the diffusion process are encoded via a fully-connected layers in the first temporal convolution of each block. The value function $V_{\boldsymbol{\psi}}$ and the barrier classifier $B_{\boldsymbol{\phi}}$ additionally have a final fully-connected layer to output a single scalar and a vector of dimension equal to the number classes (2) $\times H$ with H the planning horizon.

B. Hyperparameters

In Table 2, the hyperparameters of the experiments are reported.

Hyperparameter	Value
optimizer	AdamW (Loshchilov & Hutter, 2019)
learning rate	$2e-4$
batch size	32
training dataset size	$30K$
testing dataset size	$3K$
training epochs	200
training iterations	$[\mathbf{200K}, 300K]$
num. diffusion steps	50
planning horizon	16
CBF constant λ	0.99
value function discount γ	0.997
value guide scale η_1	$[0.1, 0.01, \mathbf{0.001}, 0.0005, 0.0001]$
classifier guide scale η_2	$[0.1, 1.0, \mathbf{5.0}, 10.0]$
tolerance ϵ	0.3
state dim.	8
action dim.	2

Table 2. DDPMs hyperparameters.

The SAC hyperparameters used are presented in Table 3.

C. Ablation Study

C.1. Guide Scale Selection

To select the value function guide scale hyperparameter η_1, we performed a grid search on a set of 100 randomly-generated targets and initial conditions using models trained $200K$ iterations. The results are reported in Table C.1

Guide Scale	Succ. Rate	Avg. Rew.	Avg. Steps
0.1	0.54	-123.81 ± 84.6	61.75 ± 39.83
0.01	0.47	-122.49 ± 88.76	63.11 ± 41.18
0.001	0.65	-104.24 ± 78.06	53.42 ± 38.42
0.0005	0.58	-104.44 ± 85.82	53.76 ± 41.94
0.0001	0.51	-125.63 ± 72.31	66.34 ± 37.14

We also performed similar experiments on a model trained for $300K$ iterations, but we did not notice substantial change in the performance.

Hyperparameter	Value
optimizer	Adam (Kingma & Ba, 2015)
learning rate	$3e-4$
batch size	32
num. training episodes	300
num. training steps	100
training dataset size	$30K$
training iterations	$[100K, 200K, 300K]$
value function discount γ	$[\mathbf{0.99}, 0.997]$
num. hidden layers	2
hidden neurons	256
hidden layers activation	ReLU
output activation	Tanh
tolerance ϵ	0.3
state dim.	8
action dim.	2
num. different seeds	3

Table 3. SAC hyperparameters.

D. Additional Experiments

Figure 4 and 5 show snapshots of a trajectory generated by following the plan generated by Algorithm 1 with value guide (see Equation (10)) and with value and safety-classifier guide (see Equation (12)). Analogously in this section, we show additional results achieved by the two different planning strategies. The code of our experiments can be found at: `github.com/nicob15`.

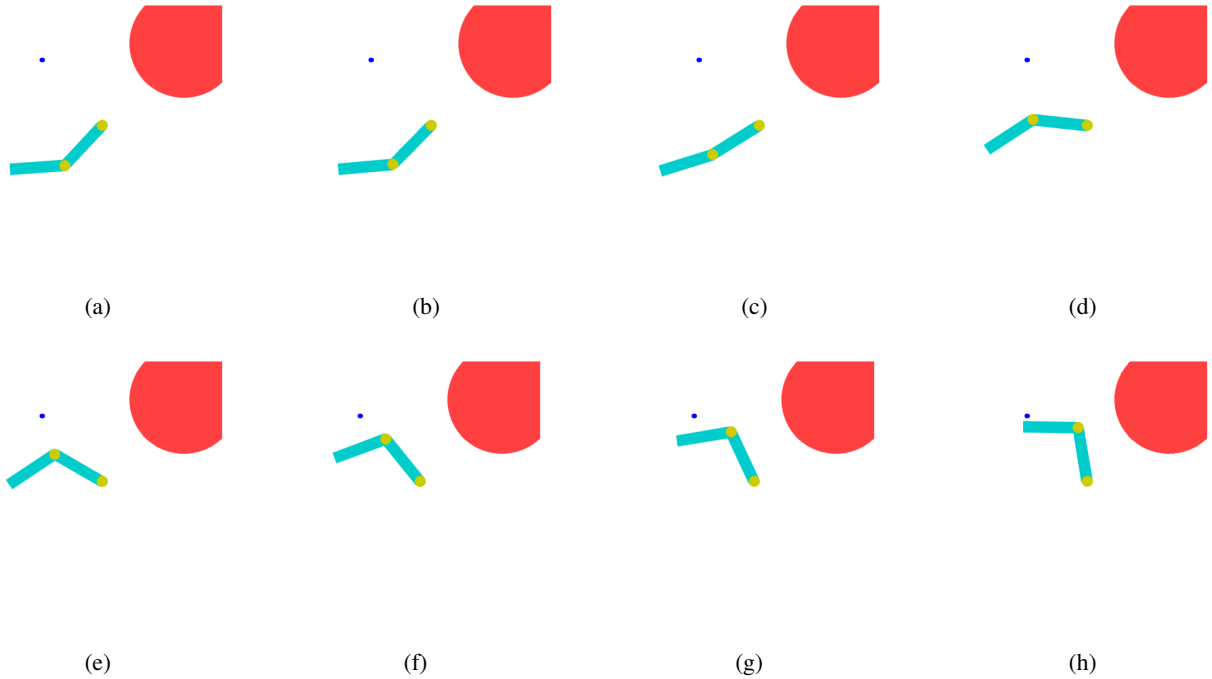


Figure 6. Trajectory generated by the diffusion planner with value-function guide.

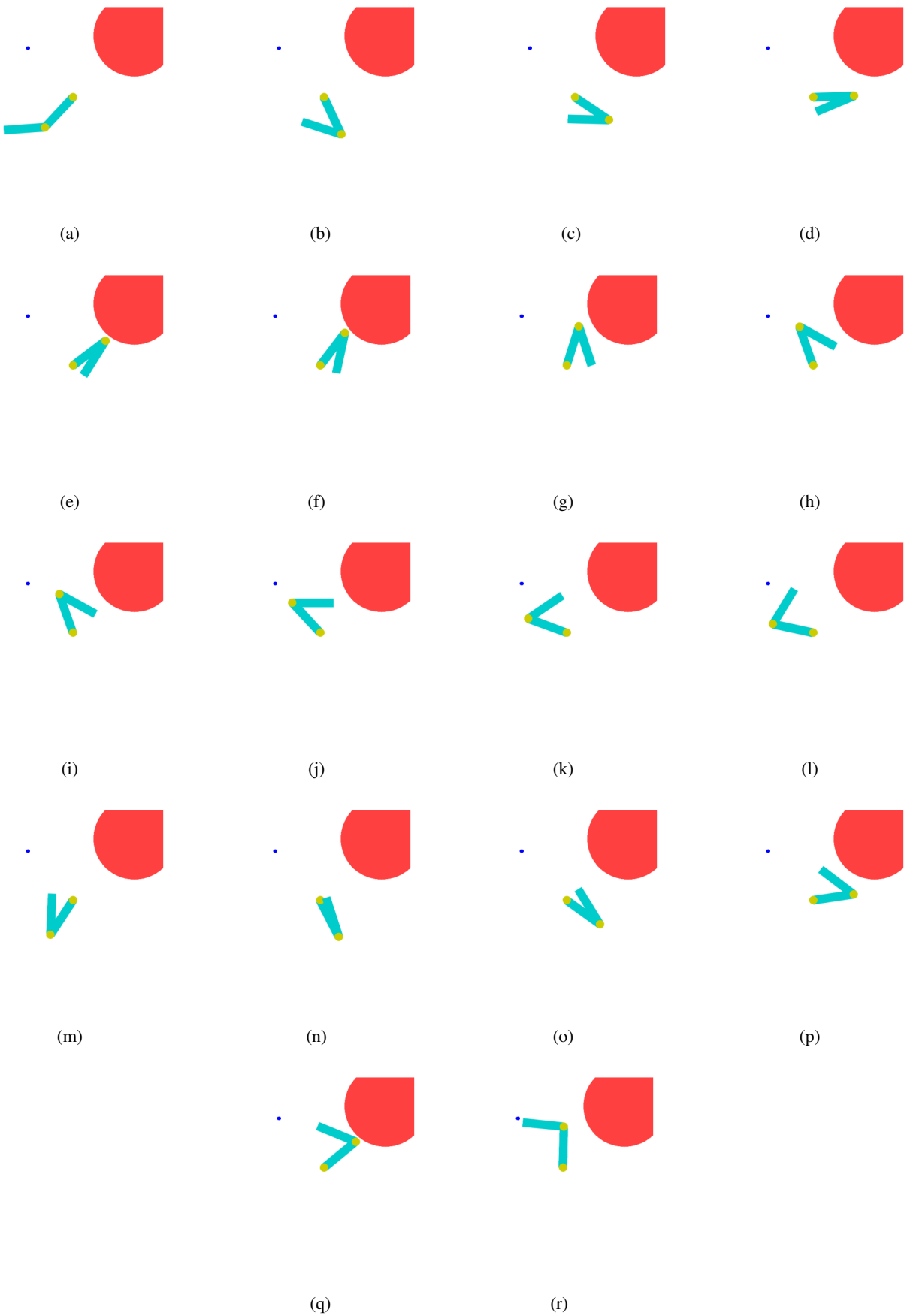


Figure 7. Trajectory generated by the diffusion planner with value-function and safety-classifier guides.

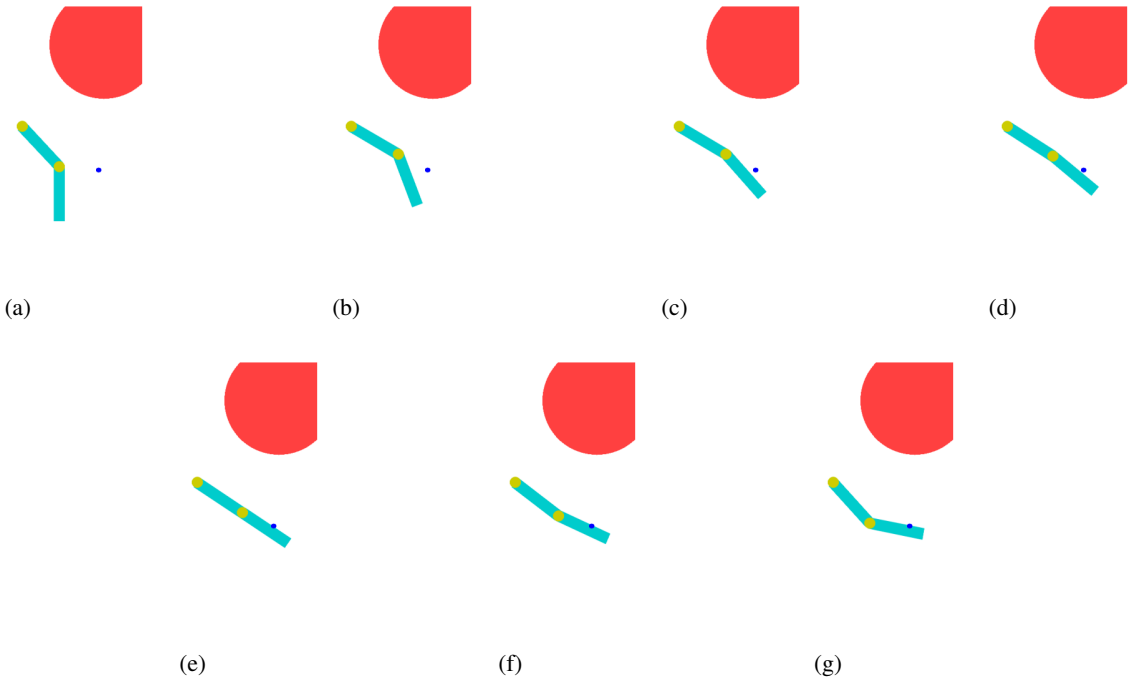


Figure 8. Trajectory generated by the diffusion planner with value-function guide.

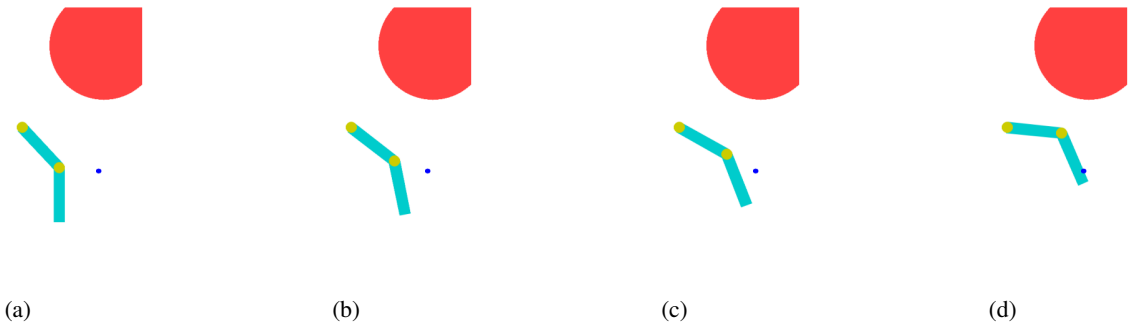
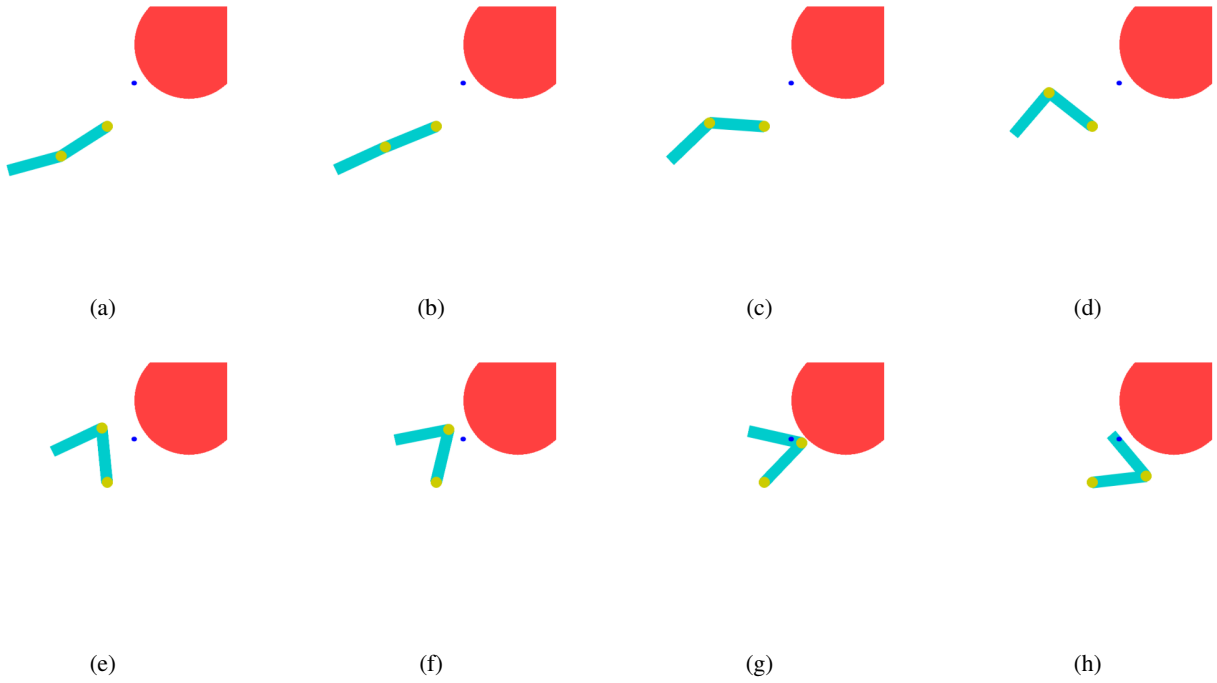


Figure 9. Trajectory generated by the diffusion planner with value-function and safety-classifier guides.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 10. Trajectory generated by the diffusion planner with value-function guide.

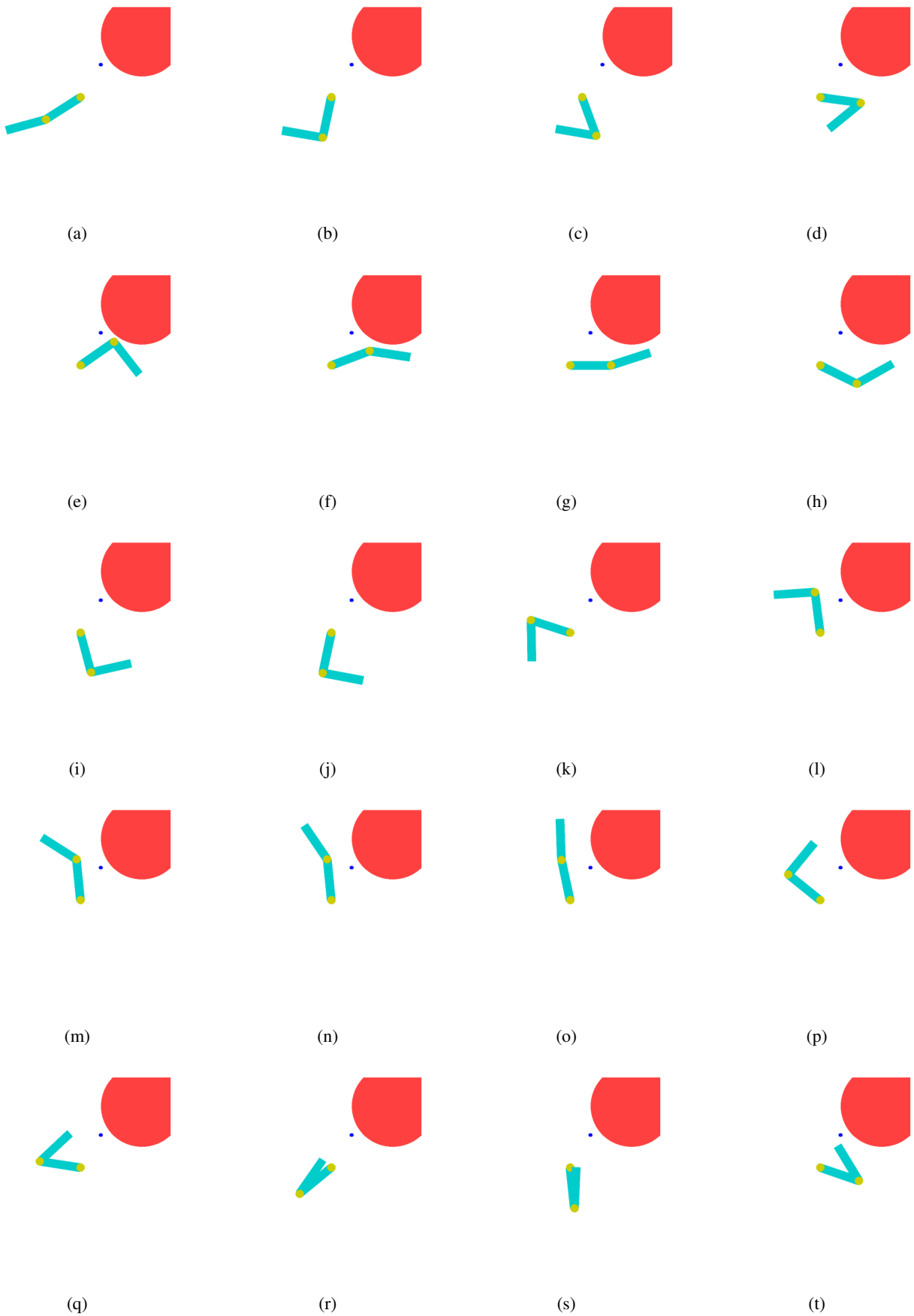


Figure 11. Trajectory generated by the diffusion planner with value-function and safety-classifier guides.

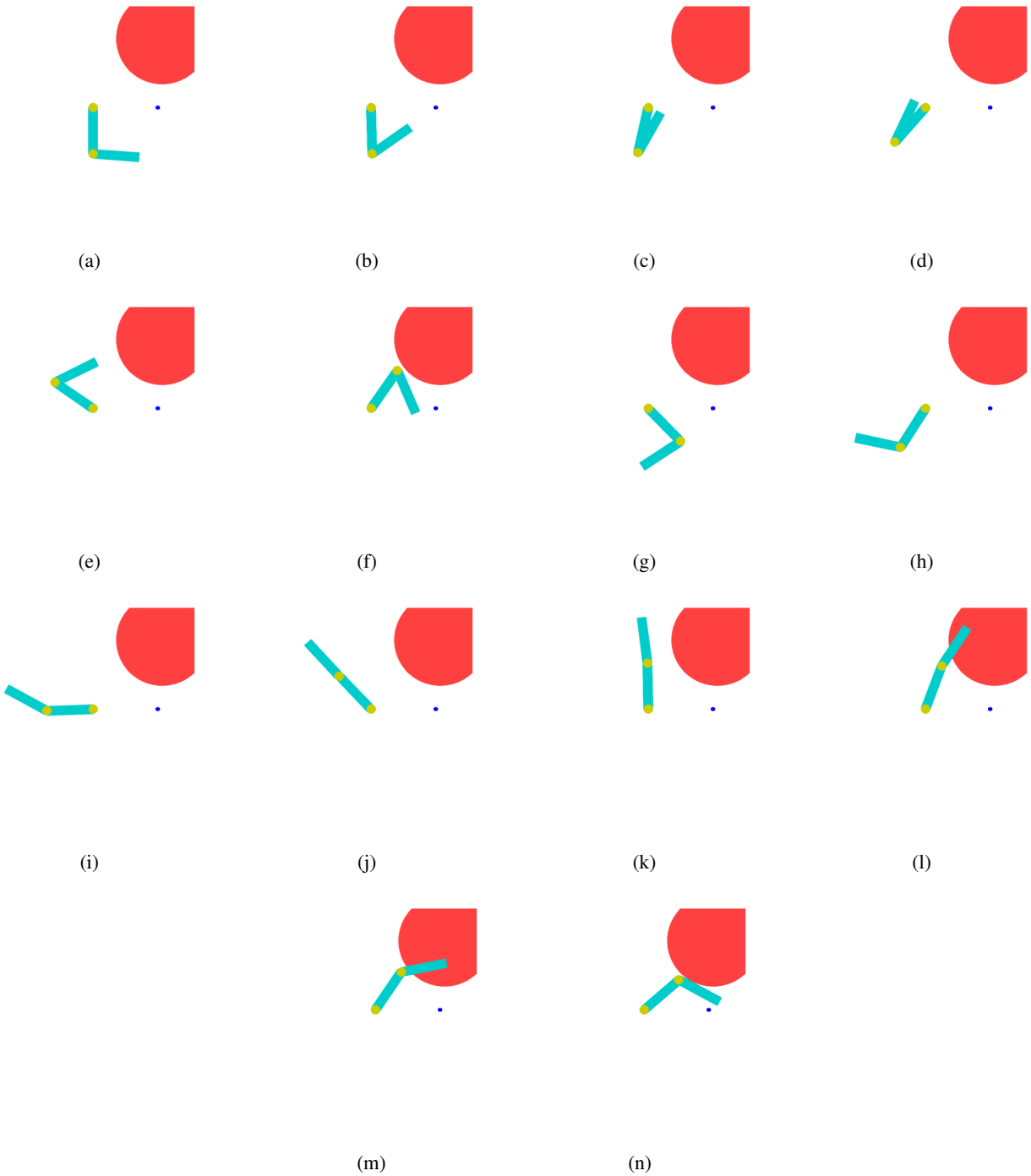


Figure 12. Trajectory generated by the diffusion planner with value-function guide.

(a) (b) (c) (d)

Figure 13. Trajectory generated by the diffusion planner with value-function and safety-classifier guides.