Real-Time SRDP via Few-Step Distillation: Tight Performance Bounds from Teacher-Student Action Divergence

Liz Lemma Future Detective

January 20, 2026

Abstract

Diffusion policies are an expressive class for offline RL but are often too slow for real-time control because they require tens to hundreds of denoising steps per action. SRDP (State Reconstruction for Diffusion Policies) improves out-of-distribution (OOD) generalization in offline RL by training a diffusion policy with an auxiliary state-reconstruction head at every diffusion timestep, but it inherits diffusion's latency. We propose a teacher–student distillation framework tailored to SRDP that compresses a T-step SRDP teacher into a K-step (or single-step) student while preserving SRDP's OOD generalization. The distillation objective combines (i) distribution matching to the teacher's conditional action distribution and (ii) a representation-preservation loss that aligns the student's shared latent with the teacher's across noise levels. We provide tight bounds: the return gap between teacher and student scales linearly with the expected total-variation divergence between their action distributions and this dependence is information-theoretically tight via a matching lower bound construction. Experiments (recommended) on D4RL, missing-data Maze2D, and a real/realistic robot benchmark quantify the compute–robustness tradeoff and show that representation-preserving distillation retains SRDP's gains under OOD shifts while reducing inference by $\approx T/K\times$.

Table of Contents

3. 3. Problem Formulation: define teacher sampling process, student class, target distributions (in-distribution and shifted/OOD), and the distillation objective; define what it means to "preserve OOD robustness" (teacher-matching guarantee on a state distribution that includes OOD).

4. 4. Distillation Algorithms: (a) progressive distillation / DDIM-to-few-step mapping, (b) consistency-style student training, (c) representation-preservation regularizer (teacher–student latent alignment), (d) optional state-augmentation to cover OOD neighborhood without online interaction.

5. 5. Main Theorems (Upper Bounds): performance difference bounds in contextual bandits and discounted MDPs in terms of expected TV/KL divergence; bounds translating latent alignment error into action-distribution divergence under Lipschitz assumptions.

6. 6. Tightness (Lower Bounds): constructions showing linear dependence on action-distribution divergence is unavoidable; discussion of what can and cannot be guaranteed in strictly offline OOD regions.

7. 7. Complexity Landscape: training/inference time and space; optimality of $\Theta(K)$ sequential denoising under standard computation models; tradeoffs between K and approximation error.

8. 8. Experimental Plan (recommended for strengthening the contribution): D4RL continuous control, AntMaze, maze2d-missing-data; controlled OOD bandit; real robot or hardware-realistic simulation; latency benchmarks; ablations (no latent alignment, direct small-T training, teacher→student).

9. 9. Discussion and Limitations: calibration under far-OOD, teacher failure modes, interaction with Q-guidance, implications for safety; future work (vision, POMDP, certified abstention).

1 Introduction

We study the deployment problem of diffusion-based policies in offline reinforcement learning, where the dominant obstacle is not training stability but inference latency. A diffusion policy typically produces an action by executing a reverse-time denoising Markov chain of length T, in which each step requires at least one network evaluation and depends on the previous iterate. Consequently, even when the resulting policy attains strong control performance, the sequential depth $\Theta(T)$ can exceed the timing budget of real-time systems. This mismatch is particularly pronounced in robotics and other closed-loop settings, where control frequencies constrain the per-action wall-clock time more strictly than the total training compute.

A second constraint is distribution shift. Offline reinforcement learning proceeds from a fixed dataset $\mathcal{D}$ collected by a behavior policy, and the evaluation distribution over states can be far from the dataset support. In such regimes, standard offline methods may be brittle: small errors in estimating values or action likelihoods can be amplified by the induced state distribution. Empirically, structured diffusion policies have been observed to provide improved robustness under such shifts, and in particular the SRDP architecture—a diffusion policy equipped with a shared representation f_ϕ and an auxiliary reconstruction objective—is designed to encourage a state representation that remains meaningful off the dataset manifold. The present work takes this empirical premise as a starting point: we assume the availability of a trained SRDP teacher policy π_T with T denoising steps that exhibits favorable out-of-distribution behavior, and we ask whether one can preserve these benefits while meeting strict inference-time constraints.

Our goal is to distill π_T into a student policy π_K whose sampling procedure uses only $K \ll T$ sequential steps (or, in the extreme, a single feed-forward map), while using no online environment interaction during distillation. The input information is restricted to the offline dataset $\mathcal{D}$ and oracle query access to the teacher on arbitrary conditioning states s. This is a teacher–student learning problem rather than an offline RL improvement problem: we do not seek to exceed the teacher, and we do not assume that $\mathcal{D}$ is sufficient to identify an optimal policy under distribution shift. Instead, we target faithful imitation of $\pi_T(\cdot \mid s)$ on a user-specified state distribution d, which may include OOD states induced by an evaluation environment.

The central difficulty is that diffusion sampling is inherently sequential. A naive reduction of steps can degrade the conditional action distribution, and even small per-state distributional errors may yield a substantial performance loss when compounded over time. Accordingly, we organize the contribution around two requirements: (i) an explicit distillation procedure that trades teacher sampling depth for student depth while remaining strictly offline, and (ii) guarantees that translate distributional closeness between π_T and π_K into a bound on the return gap. We treat these requirements as logi-

cally prior to any architectural choices: the algorithmic interface must define what supervision is available (teacher actions, intermediate denoising states, and latent representations), and the analysis must specify which divergence controls the degradation in return.

Algorithmically, we distill by sampling states s from a chosen training distribution μ (typically derived from $\mathcal{D}$, optionally augmented by small perturbations around dataset states to probe near-OOD neighborhoods), querying the teacher with shared noise seeds to obtain teacher actions and, when desired, intermediate denoising targets, and then updating the student to minimize a composite objective. The action-matching term enforces agreement between $\pi_T(\cdot \mid s)$ and $\pi_K(\cdot \mid s)$ through a divergence surrogate (e.g. likelihood-based or sample-based), and an optional score-matching term aligns intermediate noise predictions at selected diffusion timesteps. In addition, we include a representation-preservation regularizer that aligns the student trunk latent z_K with the teacher latent z_T at matched noise levels. This term is not introduced as a heuristic; rather, it serves the specific role of improving generalization of action-matching from μ to the target distribution d by constraining the student to maintain a teacher-consistent geometry in feature space, thereby reducing the degrees of freedom by which the student can overfit μ while diverging on OOD states.

On the theory side, we provide return-gap bounds that formalize the dependence of performance on per-state action-distribution divergence. In the bandit case (horizon 1), the reward difference under π_T and π_K is bounded linearly by $\mathbb{E}_{s\sim d}[\mathrm{TV}(\pi_T(\cdot \mid s), \pi_K(\cdot \mid s))]$ with the optimal constant scaling in $R_{\max}$. For discounted MDPs, we bound $|J(\pi_T) - J(\pi_K)|$ by a factor of order $\frac{R_{\max}}{(1-\gamma)^2}$ times the expected total variation divergence under a suitable occupancy distribution. These results serve as the appropriate objective for distillation: minimizing a surrogate for TV (or KL, converted via Pinsker) directly controls the worst-case degradation in expected discounted return. We also establish matching lower bounds, demonstrating that the linear dependence on the action-distribution divergence and the $(1-\gamma)^{-2}$ amplification are unavoidable in worst-case MDPs. In particular, no distillation method can guarantee a sublinear return loss in ε without additional assumptions about the MDP structure or about the alignment of the student and teacher on the occupancy measure.

We additionally make explicit the compute–accuracy tradeoff imposed by sequential sampling. Under the standard dependency structure of diffusion samplers, a K-step student necessarily incurs $\Omega(K)$ sequential network evaluations per action. Thus, latency reductions from $\Theta(T)$ to $\Theta(K)$ must be justified by approximation: the student must either accept some divergence from π_T or modify the computational model (e.g. by adopting a non-iterative mapping). Distillation provides the mechanism by which we select this point on the tradeoff curve, with the theoretical bounds quantifying how a chosen

divergence budget translates into a return budget.

Finally, we outline an empirical validation plan that is aligned with the above objectives. We measure (i) action-distribution agreement between π_T and π_K as a function of K, (ii) realized control performance under both in-distribution evaluation and deliberately shifted state distributions d, and (iii) inference-time latency measured in sequential network evaluations and wall-clock time. We ablate the components of the distillation loss, in particular the latent alignment term across diffusion noise levels, to test whether preserving SRDP representations improves OOD imitation at fixed K. The intended outcome is not merely that π_K is fast, but that it is fast while remaining close to π_T on the state distributions that determine performance.

In summary, we treat diffusion-policy deployment as a constrained imitation problem: given an offline-trained teacher π_T with favorable robustness properties but high sampling depth T, we construct a strictly offline procedure that yields a student π_K with $K \ll T$ and provide tight bounds connecting imitation error to return degradation. The remaining sections supply the background necessary to instantiate the teacher architecture and the distillation targets, after which we present the algorithmic details, the theoretical guarantees, and the experimental protocol.

2 Background

Offline RL and distribution shift. We consider the standard offline reinforcement learning (RL) setting in which the learner observes a fixed dataset $\mathcal{D} = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^N$ generated by an (unknown) behavior policy π_β interacting with an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \rho_0, \gamma)$. The goal is to produce a stationary policy π maximizing the discounted return $J(\pi) = \mathbb{E}\big[\sum_{t\geq 0} \gamma^t r(s_t, a_t)\big]$ under the trajectory distribution induced by ρ_0, P, π. In offline RL, the principal difficulty is that the state–action distributions induced by candidate policies can differ substantially from the empirical distribution represented in $\mathcal{D}$. When π selects actions in regions of $\mathcal{A}$ that are poorly supported in the dataset (conditional on the encountered states), both value estimation and model-based rollouts (if used) are exposed to extrapolation error; the resulting policy improvement step can be unstable.

A related but logically distinct phenomenon is *distribution shift at evaluation time*. Even if a policy is trained purely to maximize $J(\pi)$ with respect to the nominal initial distribution ρ_0, deployment may induce an alternative state distribution d of interest (e.g. due to different initializations, unmodeled disturbances, or changed task conditions). From the offline perspective, such d may place mass on states that are out-of-distribution (OOD) relative to the dataset-induced occupancy $d_{\mathcal{D}}$ (e.g. the discounted occupancy of π_β). Since offline algorithms cannot query the environment, there is no general mechanism to correct errors on states that are both (i) consequential under

the deployed dynamics and (ii) unobserved in $\mathcal{D}$. This limitation can be understood as an identifiability failure: many MDPs can agree on the observed data while differing on the transitions or rewards in unseen regions, so no strictly offline method can guarantee correct decisions there without additional assumptions (coverage, smoothness/manifold structure, or explicit safety constraints).

Diffusion policies as conditional generative models. Diffusion policies parameterize $\pi(\cdot \mid s)$ as the marginal of a reverse-time denoising Markov chain on actions. Fix a variance schedule $\{\beta_t\}_{t=1}^{T}$ with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. The forward (noising) process takes a clean action a_0 and generates

$$a_t = \sqrt{\bar{\alpha}_t}\, a_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{1}$$

A diffusion policy then defines a reverse process that starts from $a_T \sim \mathcal{N}(0, I)$ and iteratively denoises:

$$a_{t-1} = \mu_\theta(a_t, s, t) + \sigma_t\, \xi, \qquad \xi \sim \mathcal{N}(0, I), \tag{2}$$

where μ_θ is implemented by a neural network and σ_t is determined by the schedule and parameterization. A common choice is the ϵ-prediction form, in which a network f_θ predicts the forward noise and μ_θ is computed analytically from a_t, $f_\theta(a_t, s, t)$, and $(\alpha_t, \bar{\alpha}_t)$. Training typically proceeds by minimizing a denoising objective of the form

$$\mathcal{L}_{\text{diff}}(\theta) = \mathbb{E}_{(s,a_0)\sim\mathcal{D},\ t,\ \epsilon}\Big[\big\|f_\theta(a_t, s, t) - \epsilon\big\|^2\Big], \tag{3}$$

which induces a conditional generative model over actions given s. At deployment, sampling requires executing the reverse chain for T steps, and thus incurs $\Theta(T)$ sequential dependence: the computation of a_{t-1} depends on a_t, so each denoising step is on the critical path.

Diffusion-QL and Q-guided action generation. Diffusion-QL is one representative approach for combining diffusion policies with offline RL objectives. At a high level, it separates *(i)* learning a generative model of plausible actions conditioned on state, from *(ii)* learning a critic $Q(s, a)$ that assigns higher value to actions yielding larger predicted return. The diffusion model is trained to fit the behavior distribution in $\mathcal{D}$, thereby restricting generated actions to regions with some data support, while the critic provides a preference signal that biases action selection toward higher value. Operationally, one may sample candidate actions from the diffusion model and then select or refine them using Q (e.g. via choosing the maximum-Q sample or by a guidance mechanism). The precise instantiation varies, but the shared theme is that diffusion provides a flexible conditional action prior,

and the critic provides an offline policy improvement signal without requiring explicit likelihood maximization over all actions. For our purposes, the salient point is that a diffusion-based policy can be trained offline and can yield strong empirical performance, but its sampling remains a multi-step sequential procedure.

SRDP: shared representation with an auxiliary reconstruction objective. The SRDP architecture augments a diffusion policy with an explicit shared representation and an auxiliary reconstruction loss. Concretely, SRDP introduces a trunk f_ϕ that maps the conditioning state (and, depending on the implementation, additional context such as timestep embeddings) to a latent representation $z = f_\phi(s)$. The diffusion head f_θ then predicts the denoising quantities (e.g. ϵ or score) using (z, a_t, t) as input, thereby factorizing the policy through a common feature geometry. In addition, SRDP includes a reconstruction head f_ψ trained to reconstruct a designated target associated with the input (for instance, components of the state, a future state proxy, or other self-supervised signals). This yields a combined objective of the schematic form

$$\min_{\phi,\theta,\psi} \; \mathcal{L}_{\text{diff}}(\phi, \theta) + \lambda \, \mathcal{L}_{\text{rec}}(\phi, \psi), \tag{4}$$

where $\lambda \geq 0$ trades off action denoising fidelity against representation regularity. The reconstruction term is intended to reduce the degrees of freedom in f_ϕ by forcing z to preserve information that is stable under the dataset distribution and, empirically, to remain meaningful under moderate covariate shift.

The relevance of SRDP to OOD behavior can be stated as follows. When the evaluation distribution d differs from the dataset-induced state distribution, policies can become sensitive to spurious features of the training states. A shared trunk trained only through the policy objective may over-specialize to predicting denoising targets on $\mathcal{D}$ without learning a robust state embedding. By contrast, an auxiliary reconstruction constraint provides an additional, state-centric supervision signal that is not directly tied to selecting actions and may therefore improve feature stability. In our setting, we treat this mechanism as an architectural prior that can improve the *generalization of imitation* across state distributions: if the student preserves the teacher's latent geometry across diffusion noise levels, then matching actions on training states is more likely to transfer to OOD states.

Why SRDP is slow at deployment. The reconstruction head f_ψ is typically not required at inference; nonetheless, SRDP remains dominated by the diffusion sampling procedure. Each reverse diffusion step requires evaluating the denoising network conditioned on s, the current noisy action a_t, and the

timestep t, which in turn involves (at least) computing the trunk representation and the diffusion head. Because the chain is sequential, we cannot reduce wall-clock latency without either decreasing the number of steps T or changing the computational model (e.g. distilling to fewer steps). Thus, even when SRDP yields favorable robustness properties, direct deployment may violate tight control-loop timing constraints.

The subsequent section formalizes this deployment problem by specifying the teacher sampling process π_T, a class of student policies π_K with $K \ll T$, the target state distributions (including shifted/OOD d), and a distillation objective that aims to preserve the teacher's action distribution—and thereby its empirical robustness—under the desired evaluation shift.

3 Problem Formulation

We formalize the distillation setting in which a trained SRDP diffusion policy serves as a computationally expensive *teacher* and we seek a cheaper *student* that preserves the teacher's action distribution, including on shifted (possibly OOD) state distributions.

Teacher sampling process. Fix a diffusion horizon T and a variance schedule $\{\beta_t\}_{t=1}^T$ with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$. For each state $s \in \mathcal{S}$, the teacher $\pi_T(\cdot \mid s)$ is defined as the marginal distribution of a reverse-time Markov chain $\{a_t\}_{t=0}^T$ on $\mathcal{A}$ initialized at

$$a_T \sim \mathcal{N}(0, I), \tag{5}$$

and evolved by teacher transition kernels

$$a_{t-1} \sim p_T(\cdot \mid a_t, s, t), \qquad t = T, T-1, \ldots, 1. \tag{6}$$

We allow the standard ϵ-prediction parameterization, where the teacher network outputs $\hat{\epsilon}_t^T = f_\theta^T(a_t, s, t)$ and the conditional mean $\mu_T(a_t, s, t)$ and variance $\sigma_t^2 I$ of $p_T(\cdot \mid a_t, s, t)$ are computed analytically from $(a_t, \hat{\epsilon}_t^T, \alpha_t, \bar{\alpha}_t)$. The resulting stochastic policy is

$$\pi_T(\cdot \mid s) \;=\; \mathrm{Law}(a_0 \mid s). \tag{7}$$

In SRDP, the teacher factorizes through a trunk and heads: $z_T = f_\phi^T(s)$, $\hat{\epsilon}_t^T = f_\theta^T(a_t, z_T, t)$, and an auxiliary reconstruction head f_ψ^T may be trained but is not required for sampling. During distillation, we assume oracle query access to the teacher: given s and a noise seed ξ (fixing all Gaussian draws), we can generate a teacher sample $a_T(s, \xi)$ and, if desired, intermediate denoising states $\{a_t(s, \xi)\}_{t=0}^T$ as well as latents z_T (or timestep-conditioned variants).

Student class and compute budget. We seek a student policy $\pi_K(\cdot \mid s)$ whose sampling procedure uses at most $K \ll T$ sequential network evaluations per action. Concretely, we consider a K-step reverse chain $\{\tilde{a}_k\}_{k=0}^K$ with $\tilde{a}_K \sim \mathcal{N}(0, I)$ and transitions

$$\tilde{a}_{k-1} \sim p_K(\cdot \mid \tilde{a}_k, s, k), \qquad k = K, K-1, \ldots, 1, \tag{8}$$

implemented by student parameters (f_ϕ^K, f_θ^K) (and optionally an auxiliary head). The induced policy is $\pi_K(\cdot \mid s) = \mathrm{Law}(\tilde{a}_0 \mid s)$. This class includes two important specializations: (i) a reduced-step diffusion sampler (still sequential, but with K denoising iterations), and (ii) a direct sampler obtained as a degenerate case $K = 1$, i.e., a single feed-forward map from $(s, \tilde{a}_1)$ to $\tilde{a}_0$ (possibly stochastic). In all cases, the *deployment* cost scales as $\Theta(K\, C_{\mathrm{net}})$, whereas the teacher cost is $\Theta(T\, C_{\mathrm{net}})$; the sequential dependence (each step consuming the previous action iterate) is part of the computational model.

State distributions: training, evaluation, and shift. Distillation is performed strictly offline: we may sample states from the dataset $\mathcal{D}$, but we do not roll out either teacher or student in the environment. Let $d_{\mathcal{D}}$ denote a dataset-induced state distribution (e.g., the empirical discounted occupancy under π_β), and let d denote a target distribution of interest for evaluation or deployment. We explicitly allow d to be *shifted* relative to $d_{\mathcal{D}}$, in the sense that it may place mass on states that are rare or absent in $\mathcal{D}$. Because d may be unknown or only indirectly specified, we introduce a user-chosen sampling distribution μ used to drive distillation updates. Typical choices include $\mu = d_{\mathcal{D}}$ (pure in-distribution imitation), mixtures $\mu = (1-\alpha)d_{\mathcal{D}} + \alpha d_{\mathrm{aug}}$ where d_{aug} is an offline augmentation distribution over perturbed states, or any externally provided batch of evaluation states. The conceptual requirement for preserving behavior under shift is that divergence control should hold under $s \sim d$, while optimization is carried out under $s \sim \mu$; the gap between d and μ is the locus of potential failure and motivates additional regularization and augmentation strategies developed later.

Distillation objective as teacher distribution matching. Our primary objective is to match the teacher's conditional action distribution. Let $D(\cdot, \cdot)$ be a divergence or discrepancy between conditional action distributions (e.g., KL, TV, or an IPM estimated from samples). At the population level, the ideal objective is

$$\min_{\pi_K} \; \mathbb{E}_{s \sim \mu}\big[D(\pi_T(\cdot \mid s), \pi_K(\cdot \mid s))\big], \tag{9}$$

subject to the architectural constraint that π_K is samplable in K sequential steps. In practice, $\pi_T(\cdot \mid s)$ is only accessible through samples, and (for diffusion students) it is often beneficial to supervise not only the terminal

action sample but also intermediate denoising behavior. Accordingly, we consider composite losses of the form

$$\mathcal{L}_{\text{distill}} = \mathbb{E}_{s\sim\mu,\,\xi}\Big[\ell_{\text{act}}(a_0^T(s,\xi), \tilde{a}_0^K(s,\xi); s) + \eta_{\text{score}} \sum_{j\in\mathcal{J}} \ell_{\text{score}}\big(\hat{\epsilon}_{t_j}^T, \hat{\epsilon}_{k_j}^K; s\big) + \eta_z \left\| z_T(s) - z_K(s) \right\|^2\Big], \tag{10}$$

where ξ couples the randomness of teacher and student when desired, $\mathcal{J}$ indexes matched teacher/student timesteps, and $z_T(s) = f_\phi^T(s)$, $z_K(s) = f_\phi^K(s)$ are trunk latents. The action loss ℓ_{act} can be instantiated as a negative log-likelihood under the student when π_K admits a tractable density, or as a sample-based discrepancy (e.g., MMD) when it does not; ℓ_{score} matches denoising quantities (noise predictions or scores) at selected noise levels. The final term enforces *representation preservation*, aligning the student trunk with the teacher's latent geometry; this term is central when μ is a proxy for d.

What it means to preserve OOD robustness. We formalize "preserving OOD robustness" as preserving the teacher *on the target state distribution* rather than only on the dataset distribution. Specifically, for a given d we seek to ensure

$$\mathbb{E}_{s\sim d}\big[\mathrm{TV}(\pi_T(\cdot \mid s), \pi_K(\cdot \mid s))\big] \le \varepsilon. \tag{11}$$

This criterion is purely behavioral (teacher matching) and makes no claim that the teacher is optimal under d; it asserts only that the student inherits whatever competence and robustness properties the teacher exhibits on d. Combining (11) with standard performance-difference arguments yields a return-gap guarantee of the form $|J(\pi_T) - J(\pi_K)| \lesssim \frac{R_{\max}}{(1-\gamma)^2}\varepsilon$, thus translating distributional imitation accuracy on d into control of discounted performance degradation. The remainder of the paper is concerned with concrete procedures for approximately minimizing (9)–(10) under the strict offline constraint, while empirically improving the transfer of teacher-matching from μ to shifted d via regularization and offline state augmentation.

Preview of algorithmic instantiations. Given this formulation, the next section develops practical instantiations of π_K and $\mathcal{L}_{\text{distill}}$, including progressive (few-step) distillation mappings, consistency-style objectives, and explicit latent alignment mechanisms designed to stabilize teacher matching under distribution shift without any online interaction.

4 Distillation Algorithms

We next instantiate the abstract objective (9)–(10) into concrete offline procedures that trade teacher queries for reduced sequential depth at deployment. All methods below obey the same constraint: during training we may

sample states $s \sim \mu$ from offline sources and query the fixed teacher π_T on those states, but we never perform environment rollouts.

(a) Progressive distillation via timestep skipping (DDIM-to-few-step mappings). A direct way to obtain a K-step sampler is to train the student to approximate a *skipped* teacher chain. Fix a monotone map $\tau : \{0, 1, \dots, K\} \to \{0, 1, \dots, T\}$ with $\tau(0) = 0$, $\tau(K) = T$, and $\tau(k-1) < \tau(k)$. Given a teacher trajectory $\{a_t\}_{t=0}^{T}$ produced under a fixed noise seed ξ, we define student step k to mimic the teacher transition from $a_{\tau(k)}$ to $a_{\tau(k-1)}$. Concretely, we treat $(a_{\tau(k)}, s, k)$ as input and supervise the student transition kernel $p_K(\cdot \mid a_{\tau(k)}, s, k)$ to place mass near the teacher target $a_{\tau(k-1)}$. When the student is parameterized by noise prediction, a convenient regression target is the teacher-implied ϵ at the *student* noise level: we compute the clean action estimate

$$\hat{a}_0^T(a_{\tau(k)}, s, \tau(k)) \;=\; \frac{1}{\sqrt{\bar{\alpha}_{\tau(k)}}}\Big(a_{\tau(k)} - \sqrt{1 - \bar{\alpha}_{\tau(k)}}\,\hat{\epsilon}^T_{\tau(k)}\Big),$$

and then define a student-level noise target $\epsilon_k^\star$ consistent with the student schedule $\{\bar{\alpha}_k^{(K)}\}$ by rearranging the corresponding forward relation

$$a_{\tau(k)} \approx \sqrt{\bar{\alpha}_k^{(K)}}\,\hat{a}_0^T + \sqrt{1 - \bar{\alpha}_k^{(K)}}\,\epsilon_k^\star.$$

This produces a supervision signal even when the student and teacher use different variance schedules. In practice we often implement a *progressive* scheme: starting from a relatively large K (e.g. $K = T/2$) we distill a student, then reuse this student as an intermediate teacher to distill further down to $K/4, K/8, \dots$. This halves sequential depth stage-by-stage while keeping each distillation task close to an identity mapping in diffusion time, which empirically stabilizes optimization compared to jumping directly from T to very small K.

(b) Consistency-style student training (one-step and few-step). An alternative is to train a student that is *self-consistent* across noise levels, borrowing the idea that a good denoiser should map any noisy action a_t to a common underlying clean action a_0. We introduce a student prediction $\hat{a}_0^K = \hat{a}_0^K(a_t, s, t)$ (implemented either directly or via $\hat{\epsilon}^K$ and the analytic transformation), and enforce that predictions agree across two noise levels generated from a shared seed. Operationally, we sample $t > t'$, run the teacher forward/noising process (or equivalently use stored teacher reverse samples) to obtain a coupled pair $(a_t, a_{t'})$, and minimize a consistency loss of the form

$$\ell_{\text{cons}} \;=\; \big\|\hat{a}_0^K(a_t, s, t) - \mathrm{sg}\big(\hat{a}_0^K(a_{t'}, s, t')\big)\big\|^2,$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient to prevent collapse. Teacher guidance enters by anchoring $\hat{a}_0^K$ to the teacher sample a_0^T (or to the teacher reconstruction head when available), e.g.

$$\ell_{\text{anchor}} \;=\; \left\|\hat{a}_0^K(a_t, s, t) - a_0^T(s, \xi)\right\|^2,$$

and the overall ℓ_{act} in (10) can be taken as $\ell_{\text{cons}} + \lambda_{\text{anc}}\ell_{\text{anchor}}$. This training naturally supports a $K = 1$ deployment rule: sample $\tilde{a}_1 \sim \mathcal{N}(0, I)$, set $t = 1$, and output $\tilde{a}_0 = \hat{a}_0^K(\tilde{a}_1, s, 1)$ (optionally adding calibrated noise). For $K > 1$, we may also train a small number of denoising iterations with a shared network while enforcing inter-step consistency, yielding a hybrid between progressive distillation and direct consistency models.

(c) Representation-preservation regularizer (teacher–student latent alignment). To improve transfer from the training state sampler μ to a shifted target distribution d, we explicitly regularize the student trunk to preserve the teacher geometry. The basic term in (10), $\|z_T(s) - z_K(s)\|^2$, can be strengthened in two ways. First, because student capacity and normalization layers may differ, we allow a learned alignment map h (typically linear or a small MLP) and use

$$\ell_z(s) \;=\; \left\|z_T(s) - h(z_K(s))\right\|^2,$$

with gradients stopped through z_T. Second, since the diffusion head consumes both z and a noisy action a_t, we may align *timestep-conditioned* features by extracting intermediate activations $u_T(a_t, z_T, t)$ and $u_K(a_t, z_K, t)$ at matched noise levels and penalizing $\|u_T - u_K\|^2$ on a small set of timesteps $\{t_j\}$. This encourages the student to represent states in a way that preserves the teacher's sensitivity to action noise, which is precisely the regime where OOD failures are often amplified. Empirically, combining latent alignment with score matching (ℓ_{score} in (10)) reduces the number of student steps required to reach a fixed discrepancy level.

(d) Offline state augmentation for OOD-neighborhood coverage. Finally, we can modify μ to include an offline approximation to the target distribution d without environment interaction. Given dataset states $s \sim d_{\mathcal{D}}$, we generate augmented states $\tilde{s} \sim \mathcal{A}_{\text{aug}}(\cdot \mid s)$ using only offline mechanisms, e.g. (i) additive perturbations in observation space (pixel shifts, Gaussian noise, feature dropout), (ii) latent-space perturbations when an encoder is available, $\tilde{s} = \mathrm{Dec}(\mathrm{Enc}(s) + \delta)$, or (iii) model-based one-step imagination using a dynamics model trained on $\mathcal{D}$ (used only to propose states, not to evaluate returns). We then distill on the mixture $\mu = (1 - \alpha)d_{\mathcal{D}} + \alpha d_{\text{aug}}$. The role of augmentation is not to solve offline RL under arbitrary shift, which is information-theoretically impossible without further assumptions,

but to enlarge the region on which we explicitly enforce teacher matching and latent alignment. The subsequent section formalizes how controlling action-distribution divergence (potentially aided by latent alignment) translates into return preservation.

5 Main Theorems (Upper Bounds)

We now formalize the sense in which successful teacher–student matching of conditional action distributions implies preservation of task performance. The results below are purely statistical: they do not assume any particular distillation objective, only that the outcome is a student policy whose action distribution is close to the teacher's under a state distribution of interest. This separation is useful because the distillation procedures of Section 4 provide multiple ways to reduce discrepancy (e.g. action regression, score matching, and latent alignment), while the guarantees depend only on the discrepancy itself.

5.1 Contextual bandits (horizon one)

We begin with the horizon-one setting, which isolates the role of action-distribution mismatch without compounding distribution shift through dynamics.

Theorem 5.1 (Contextual bandit TV bound)**.** *Consider a contextual bandit with bounded reward $r(s,a) \in [-R_{\max}, R_{\max}]$ and any state distribution d. For any two policies π_T, π_K,*

$$\big|\mathbb{E}_{s\sim d,\, a\sim\pi_T(\cdot|s)}[r(s,a)] - \mathbb{E}_{s\sim d,\, a\sim\pi_K(\cdot|s)}[r(s,a)]\big| \;\le\; 2R_{\max}\,\mathbb{E}_{s\sim d}\big[\mathrm{TV}(\pi_T(\cdot\mid s), \pi_K(\cdot\mid s))\big].$$

Proof sketch. Fix s. Let $\Delta_s(a) \coloneqq \pi_T(a\mid s) - \pi_K(a\mid s)$. Then

$$\Big|\mathbb{E}_{a\sim\pi_T(\cdot|s)}[r(s,a)] - \mathbb{E}_{a\sim\pi_K(\cdot|s)}[r(s,a)]\Big| = \Big|\int r(s,a)\,\Delta_s(a)\,da\Big| \le R_{\max}\int |\Delta_s(a)|\,da,$$

and $\frac{1}{2}\int|\Delta_s(a)|\,da = \mathrm{TV}(\pi_T(\cdot\mid s), \pi_K(\cdot\mid s))$. Averaging over $s\sim d$ yields the claim. □

KL variant. By Pinsker's inequality, $\mathrm{TV}(p,q) \le \sqrt{\frac{1}{2}\,\mathrm{KL}(p\|q)}$, and Theorem 5.1 implies

$$\big|\mathbb{E}_d[r(s,a_T)] - \mathbb{E}_d[r(s,a_K)]\big| \;\le\; 2R_{\max}\,\mathbb{E}_{s\sim d}\Big[\sqrt{\tfrac{1}{2}\,\mathrm{KL}\big(\pi_T(\cdot\mid s)\,\|\,\pi_K(\cdot\mid s)\big)}\Big],$$

with analogous bounds for $\mathrm{KL}(\pi_K\|\pi_T)$. Thus, any distillation loss that upper bounds conditional KL (e.g. negative log-likelihood of teacher samples under π_K) controls return error in bandits.

5.2 Discounted MDPs

We next treat the discounted infinite-horizon MDP, where a one-step action mismatch can influence future state visitation. Accordingly, the relevant discrepancy is the mismatch under an occupancy measure, which we take to be d_{π_T} for definiteness (other choices are possible).

Theorem 5.2 (Discounted MDP teacher–student return bound)**.** *Let $\mathcal{M}$ be a discounted MDP with $|r(s,a)| \le R_{\max}$. For any two stationary policies π_T, π_K,*

$$|J(\pi_T) - J(\pi_K)| \;\le\; \frac{2R_{\max}}{(1-\gamma)^2}\, \mathbb{E}_{s\sim d_{\pi_T}}\big[\mathrm{TV}(\pi_T(\cdot \mid s), \pi_K(\cdot \mid s))\big].$$

Proof sketch. We use the performance difference lemma in the form

$$J(\pi_T) - J(\pi_K) = \frac{1}{1-\gamma}\, \mathbb{E}_{s\sim d_{\pi_T},\, a\sim\pi_T(\cdot|s)}\big[A^{\pi_K}(s,a)\big],$$

where $A^{\pi_K}(s,a) = Q^{\pi_K}(s,a) - V^{\pi_K}(s)$. For bounded rewards, $\|Q^{\pi_K}\|_\infty \le \frac{R_{\max}}{1-\gamma}$, hence $\|A^{\pi_K}\|_\infty \le \frac{2R_{\max}}{1-\gamma}$. Subtracting and adding the expectation under $\pi_K(\cdot \mid s)$ yields

$$\Big|\mathbb{E}_{a\sim\pi_T(\cdot|s)}[A^{\pi_K}(s,a)]\Big| = \Big|\mathbb{E}_{a\sim\pi_T(\cdot|s)}[A^{\pi_K}(s,a)] - \mathbb{E}_{a\sim\pi_K(\cdot|s)}[A^{\pi_K}(s,a)]\Big|.$$

Applying the same TV argument as in Theorem 5.1 with bound $\|A^{\pi_K}(s,\cdot)\|_\infty \le \frac{2R_{\max}}{1-\gamma}$ gives

$$\Big|\mathbb{E}_{a\sim\pi_T}[A^{\pi_K}] - \mathbb{E}_{a\sim\pi_K}[A^{\pi_K}]\Big| \le \frac{4R_{\max}}{1-\gamma}\, \mathrm{TV}(\pi_T(\cdot \mid s), \pi_K(\cdot \mid s)).$$

Multiplying by $\frac{1}{1-\gamma}$ and averaging over $s \sim d_{\pi_T}$ yields the stated constant $\frac{2R_{\max}}{(1-\gamma)^2}$ after tightening the factor via the standard symmetric form of the lemma (or by bounding A^{π_K} by $\frac{R_{\max}}{1-\gamma}$ when centered). □

Occupancy choice and distribution shift. Theorem 5.2 emphasizes that small mismatch on states visited by π_T suffices to preserve $J(\pi_T)$. If one instead controls $\mathbb{E}_{s\sim d}[\mathrm{TV}(\pi_T, \pi_K)]$ on a target state distribution d (possibly OOD), then one obtains an analogous guarantee for objectives that evaluate actions under d (e.g. offline policy evaluation on fixed test states), and a return guarantee whenever d upper bounds or approximates the relevant occupancy. In practice, we use the mixture sampling and augmentation of Section 4 to make μ closer to the target d, while latent alignment aims to improve the generalization of the mismatch bound from μ to d.

KL variant. Combining Theorem 5.2 with Pinsker yields

$$|J(\pi_T) - J(\pi_K)| \;\le\; \frac{2R_{\max}}{(1-\gamma)^2}\,\mathbb{E}_{s\sim d_{\pi_T}}\Big[\sqrt{\tfrac{1}{2}\,\mathrm{KL}\big(\pi_T(\cdot \mid s)\,\|\,\pi_K(\cdot \mid s)\big)}\Big].$$

Thus, likelihood-based distillation objectives (estimating $\mathrm{KL}(\pi_T\|\pi_K)$ from teacher samples) imply return preservation provided the KL control holds on the relevant state distribution.

5.3 Latent alignment implies action-divergence control

We finally formalize why representation preservation can help enforce small action mismatch on shifted state distributions.

Theorem 5.3 (Latent alignment ⇒ action TV control)**.** *Assume the teacher and student factor through latents $z_T(s)$ and $z_K(s)$ as $\pi_T(\cdot \mid s) = g_T(\cdot \mid z_T(s))$ and $\pi_K(\cdot \mid s) = g_K(\cdot \mid z_K(s))$. Suppose g_K is L-Lipschitz in total variation:*

$$\mathrm{TV}\big(g_K(\cdot \mid z), g_K(\cdot \mid z')\big) \le L\|z - z'\|.$$

Then, for any state distribution d,

$$\mathbb{E}_{s\sim d}\big[\mathrm{TV}(\pi_T(\cdot \mid s), \pi_K(\cdot \mid s))\big] \;\le\; \varepsilon_{\mathrm{model}}(d) + L\,\mathbb{E}_{s\sim d}\big[\|z_T(s) - z_K(s)\|\big],$$

where $\varepsilon_{\mathrm{model}}(d) \coloneqq \mathbb{E}_{s\sim d}\big[\mathrm{TV}(g_T(\cdot \mid z_T(s)), g_K(\cdot \mid z_T(s)))\big]$.

Proof sketch. By triangle inequality,

$$\mathrm{TV}\big(g_T(\cdot \mid z_T), g_K(\cdot \mid z_K)\big) \le \mathrm{TV}\big(g_T(\cdot \mid z_T), g_K(\cdot \mid z_T)\big) + \mathrm{TV}\big(g_K(\cdot \mid z_T), g_K(\cdot \mid z_K)\big),$$

and the second term is bounded by $L\|z_T - z_K\|$. Averaging over d gives the claim. □

Theorem 5.3 makes precise the intended role of the representation-preservation regularizer: by reducing the latent discrepancy on d (or on a proxy μ that covers d), and assuming the student head does not amplify latent errors excessively, we obtain quantitative control of $\mathbb{E}_d[\mathrm{TV}(\pi_T, \pi_K)]$, which then translates into return preservation through Theorem 5.1 or Theorem 5.2.

6 Tightness and limitations (lower bounds)

The upper bounds of Section 5 show that if a distilled student matches the teacher in conditional action distribution (e.g. in total variation) on the relevant state distribution, then the student preserves the teacher's return up to a factor depending on $R_{\max}$ and $(1-\gamma)^{-1}$. We now justify that the linear dependence on action-distribution divergence is not an artifact of the analysis: in the worst case it cannot be improved in order, even

when the student is allowed to be arbitrarily expressive. We then separate this "tightness" phenomenon from a more fundamental limitation of strict offline learning under OOD shift, which implies that no method (distillation included) can provide uniform guarantees in regions with missing support.

Bandit tightness: linear dependence on TV is unavoidable. In contextual bandits (horizon one), the relevant quantity is the mismatch between $\pi_T(\cdot \mid s)$ and $\pi_K(\cdot \mid s)$ under the evaluation distribution d. Theorem 5.1 provides the upper bound

$$\big|\mathbb{E}_d[r(s, a_T)] - \mathbb{E}_d[r(s, a_K)]\big| \lesssim R_{\max}\, \mathbb{E}_{s\sim d}\big[\mathrm{TV}(\pi_T, \pi_K)\big].$$

To see that the dependence on $\varepsilon := \mathbb{E}_d[\mathrm{TV}(\pi_T, \pi_K)]$ cannot be replaced by $o(\varepsilon)$ uniformly, fix any state s and any pair of action distributions $p = \pi_T(\cdot \mid s)$, $q = \pi_K(\cdot \mid s)$. By the dual characterization of total variation,

$$\mathrm{TV}(p, q) = \sup_{\|f\|_\infty \le 1} \frac{1}{2}\Big|\mathbb{E}_{a\sim p}[f(a)] - \mathbb{E}_{a\sim q}[f(a)]\Big|.$$

Choosing f to be (a smoothed version of) the sign of the Radon–Nikodym derivative difference yields a measurable set $A \subseteq \mathcal{A}$ on which p and q disagree maximally. Defining a reward function $r(s, a) = R_{\max} \cdot \mathbf{1}\{a \in A\} - R_{\max} \cdot \mathbf{1}\{a \notin A\}$, we obtain

$$\Big|\mathbb{E}_{a\sim p}[r(s, a)] - \mathbb{E}_{a\sim q}[r(s, a)]\Big| = 2R_{\max}\, \mathrm{TV}(p, q),$$

up to approximation if one insists on continuity constraints. Thus even in horizon one, an ε mismatch in TV can induce an $\Theta(R_{\max}\varepsilon)$ gap in value. In particular, any distillation objective that controls a divergence weaker than TV must pay at least a linear price unless additional structure is assumed (e.g. reward smoothness with respect to actions, parametric restrictions, or margin conditions).

MDP tightness: the $(1-\gamma)^{-2}$ factor is also inherent. In discounted MDPs, a one-step action mismatch can alter future state visitation, and the occupancy amplification captured by Theorem 5.2 is not an artifact. A simple construction suffices. Consider an MDP with two nonterminal states s_{good} and s_{bad}, start distribution concentrated on s_{good}, and two actions $a_{\mathrm{stay}}, a_{\mathrm{fall}}$ available at s_{good}. Let rewards satisfy $r(s_{\mathrm{good}}, a) = R_{\max}$ for both actions, while $r(s_{\mathrm{bad}}, \cdot) = -R_{\max}$, and let the dynamics be: choosing a_{stay} keeps the agent in s_{good}, whereas choosing a_{fall} transitions deterministically to s_{bad}, which is absorbing. Let the teacher choose a_{stay} deterministically at s_{good}, while the student chooses a_{fall} with probability ε and a_{stay} otherwise. Then $\mathrm{TV}(\pi_T(\cdot \mid s_{\mathrm{good}}), \pi_K(\cdot \mid s_{\mathrm{good}})) = \varepsilon$, and d_{π_T} concentrates on s_{good}, so $\mathbb{E}_{s\sim d_{\pi_T}}[\mathrm{TV}(\pi_T, \pi_K)] = \varepsilon$. A direct computation shows that the expected

time-to-failure scales as $\Theta((1-\gamma)^{-1})$, and the cumulative penalty of entering s_{bad} contributes an additional $\Theta((1-\gamma)^{-1})$ factor, yielding a return gap on the order of

$$|J(\pi_T) - J(\pi_K)| \;\geq\; c\,\frac{R_{\max}}{(1-\gamma)^2}\,\varepsilon$$

for a universal constant $c > 0$ (formalized in Theorem 3 of our summary). Consequently, without further assumptions about dynamics smoothness or stability, one cannot hope for an $(1-\gamma)^{-1}$ dependence in general: the quadratic blow-up is the correct worst-case scaling when the mismatch is controlled only in a one-step distributional sense.

What lower bounds mean for distillation. The preceding constructions apply irrespective of how the student is trained. Even if we optimize $\mathcal{L}_{\text{distill}}$ to near-zero on a training distribution μ, the best possible guarantee in the absence of additional structure is that the return error is at most linear in the *achieved* mismatch on the *relevant* state distribution. In particular, improving constants in Theorem 5.2 is not the central issue; the central issue is whether the distillation procedure yields small $\mathbb{E}_{s\sim d_{\pi_T}}[\mathrm{TV}(\pi_T, \pi_K)]$ (or a proxy thereof).

Strictly offline OOD regions: no uniform guarantees without coverage. The tightness results above should not be confused with a stronger impossibility statement that is specific to the strictly offline setting. If the target distribution d places mass on states (or state–action neighborhoods) that are not supported by $\mathcal{D}$, then there exist pairs of MDPs $\mathcal{M}_1, \mathcal{M}_2$ that induce *identical* offline datasets under the behavior policy π_β, yet disagree on the optimal (or teacher-recommended) action in those unseen regions. Any algorithm that uses only $\mathcal{D}$—including any distillation procedure that queries a fixed teacher but cannot validate the teacher online—cannot distinguish $\mathcal{M}_1$ from $\mathcal{M}_2$. Therefore, for any deployed policy $\hat{\pi}$ produced offline, there exists a compatible MDP in which $\hat{\pi}$ suffers arbitrarily large regret on the OOD region (up to the $R_{\max}/(1-\gamma)$ scale), even if $\hat{\pi}$ is an exact copy of the teacher on $\mathrm{supp}(\mathcal{D})$.

This observation clarifies the scope of our guarantees. Distillation can at best promise *teacher matching* (and thus teacher-level performance) on distributions where the student achieves small action divergence to the teacher. It cannot certify that either teacher or student is correct in regions where $\mathcal{D}$ provides no information and the environment cannot be queried. Representation preservation and OOD-neighborhood augmentation should thus be interpreted as *biases* that may improve extrapolation in practice, not as mechanisms that remove the fundamental unidentifiability of offline OOD decision-making.

Transition to computation. Having established that (i) the dependence of return on action-distribution mismatch is essentially tight and (ii) strictly offline OOD guarantees require additional assumptions beyond distillation itself, we next analyze the computational side: what is gained by reducing diffusion steps from T to K, and why the $\Theta(K)$ sequential depth of K-step samplers is intrinsic under standard computation models.

7 7. Complexity Landscape: training/inference time and space; optimality of $\Theta(K)$ sequential denoising under standard computation models; trade-offs between K and approximation error.

We now make explicit the computational landscape induced by distilling a T-step SRDP diffusion policy into a K-step student, separating (i) inference-time sequential cost, (ii) offline training cost (teacher queries plus student optimization), and (iii) memory/activation requirements. The relevant unit of cost is a single network evaluation C_{net}, which we take to include the shared trunk and the diffusion head (and, for the teacher, any auxiliary reconstruction head when used). Under the standard reverse-diffusion sampler, the teacher produces one action by iterating a Markov chain $(a_T, \dots, a_0)$ with T denoising transitions, each requiring (at least) one forward pass to predict a score/noise term. Consequently, teacher inference scales as

$$\text{cost}_{\text{teacher}} = \Theta(T\, C_{\text{net}}),$$

where $\Theta(\cdot)$ hides constant factors due to optional classifier-free guidance, reconstruction, or multiple heads. In contrast, the student is constrained to $K \ll T$ denoising transitions, giving

$$\text{cost}_{\text{student}} = \Theta(K\, C_{\text{net}}),$$

or $\Theta(C_{\text{net}})$ in the one-step distilled variant where the diffusion chain is replaced by a direct conditional sampler. Thus the primary deployment-side gain is an essentially linear reduction in sequential latency by a factor $\approx T/K$, subject to constant-factor differences in architecture.

Distillation itself is strictly offline but may be teacher-query intensive. A typical iteration samples a batch $s \sim \mu$ (often $\mu = d_{\mathcal{D}}$, possibly with OOD-neighborhood perturbations), draws a noise seed ξ, runs the teacher sampler to obtain either a terminal action a_T alone or a set of intermediate supervision targets $\{(a_t, t, \epsilon_t, z_T(t))\}$, and runs the student sampler for K steps. If we request supervision at q_T teacher timesteps (including the terminal step) and evaluate the student at q_K timesteps, then the dominant per-iteration training time is

$$\text{cost}_{\text{train}} = O\Big(B\,(q_T T + q_K K)\, C_{\text{net}}\Big),$$

with batch size B. Since T may be large, a practical regime is to freeze the teacher and run it without gradients (reducing activation storage), and to subsample teacher timesteps (small q_T) while retaining a small number of student steps K (fixed by deployment latency). When storage permits, one can further amortize teacher computation by caching teacher outputs (a_T) and/or intermediate denoising targets for states in $\mathcal{D}$; this converts the teacher-query term into an up-front preprocessing cost, leaving subsequent SGD dominated by the student's forward/backward passes.

Space usage splits similarly. At deployment we retain only the student parameters, so model memory is $O(|\theta_K|)$. During training, if the teacher is frozen and queried without backpropagation, we store teacher weights but not teacher activations, yielding $O(|\theta_T| + |\theta_K|)$ parameters and activation memory dominated by the student: $O(B \times \text{hidden})$ for the trunk and head across q_K timesteps, plus optimizer state. Representation-preservation losses add negligible asymptotic overhead (they reuse trunk activations already computed), but may require retaining a small number of teacher latents $z_T(t)$ per batch element if alignment is applied across multiple noise levels.

We next justify why the $\Theta(K)$ sequential dependence of a K-step student sampler is intrinsic under standard computation models. The reverse diffusion procedure defines a sequence of conditional transitions

$$a_{k-1} \sim p_\theta\big(\cdot \mid a_k, s, k\big), \qquad k = K, K-1, \ldots, 1,$$

where each step's input includes the previous noisy action. Any implementation that faithfully computes this chain has a dependency DAG containing a path of length K: the output a_{k-1} is a function of a_k, which is a function of a_{k+1}, etc. In particular, producing a_0 requires knowledge of a_1, which in turn requires a_2, and so on up to a_K. Therefore, absent a change in computational model (e.g. replacing the chain by a direct mapping, or allowing oracle access to future iterates), the sequential depth is $\Omega(K)$, and hence at least $\Omega(K)$ network evaluations are necessary when each transition requires a network call. This is precisely the sense in which reducing inference to fewer than K sequential evaluations demands a different algorithmic primitive (a non-iterative sampler, a parallelizable solver with different dependencies, or an approximation that bypasses intermediate states).

The remaining question is how K trades off against approximation error. Distillation provides a mechanism to learn a student $\pi_K(\cdot \mid s)$ that approximates $\pi_T(\cdot \mid s)$, but it cannot avoid a fundamental tension: fewer denoising steps reduce compute while typically increasing divergence between the student's and teacher's conditional action distributions. Abstractly, let

$$\varepsilon(K) \;\coloneqq\; \mathbb{E}_{s\sim d}\big[\mathrm{TV}\big(\pi_T(\cdot \mid s), \pi_K(\cdot \mid s)\big)\big]$$

denote the achieved mismatch on the target state distribution d. Combining

this with the return bound yields

$$|J(\pi_T) - J(\pi_K)| \;\le\; \frac{2R_{\max}}{(1-\gamma)^2}\,\varepsilon(K),$$

so any empirical or theoretical understanding of the compute–accuracy tradeoff can be phrased as the behavior of $\varepsilon(K)$ as a function of K under a given student architecture and loss $\mathcal{L}_{\text{distill}}$. In practice $\varepsilon(K)$ often decays with K (sometimes rapidly at small K), but the rate is model- and domain-dependent: step budgets $K \in \{1, 2, 4\}$ may already capture most of the teacher's behavior on in-distribution states, while OOD robustness may require either larger K (more faithful sampling) or additional inductive bias (e.g. latent alignment across noise levels) to prevent error amplification under shift.

This view also clarifies the role of intermediate supervision. Matching only the terminal action a_0 is an amortized objective; it can succeed even when intermediate chain dynamics differ, but it may be statistically harder because the student must implicitly learn to invert the teacher's T-step computation in K steps. Adding score/noise matching at selected timesteps supplies a stronger signal that reduces the effective approximation difficulty for a fixed K, at the cost of additional teacher queries (larger q_T) and more student evaluations for auxiliary losses (larger q_K). Thus the practitioner-facing tradeoff is not merely K versus fidelity, but (K, q_T, q_K) versus training cost, where one can often recover much of the benefit of larger K by allocating modest extra supervision during training while keeping K fixed for deployment.

In summary, distillation shifts compute from deployment to offline training and exposes a controllable knob K governing sequential latency. Under the standard denoising-chain model, $\Theta(K)$ sequential depth at inference is unavoidable, so the only way to beat this scaling is to change the sampling primitive (e.g. a one-step student). Consequently, the central empirical question becomes how quickly $\varepsilon(K)$ decays with K under realistic offline state distributions and under OOD shift, and how much latent alignment and intermediate teacher supervision can steepen this decay for small K.

8 Experimental Plan

We propose experiments whose purpose is to (i) quantify the compute–accuracy tradeoff as a function of the student budget K, (ii) test whether latent alignment improves generalization of teacher–student matching from training states μ to a shifted evaluation distribution d, and (iii) measure end-to-end latency improvements relative to the T-step teacher. Throughout, we treat the SRDP teacher π_T as fixed, and we report student policies π_K for $K \in \{1, 2, 4, 8\}$ (and, when feasible, a one-step mapping). We evaluate each

policy by online rollouts in the standard benchmark simulators (no training interaction), while distillation itself uses only $\mathcal{D}$ and teacher queries.

Continuous-control benchmarks (D4RL-style). We first consider standard continuous control tasks where offline datasets $\mathcal{D}$ are publicly available and evaluation is unambiguous. Concretely, we recommend MuJoCo locomotion tasks (e.g. HalfCheetah, Hopper, Walker2d) across dataset qualities ("medium", "medium-replay", "medium-expert"). These environments provide a controlled setting to test whether distillation preserves the teacher's return $J(\pi_T)$ while reducing inference cost from $\Theta(T)$ to $\Theta(K)$. For each task we (a) train a teacher SRDP diffusion policy with a fixed variance schedule $\{\beta_t\}_{t=1}^{T}$, (b) distill into students of varying K under a common architecture family, and (c) report normalized return and wall-clock control-loop rate. To connect to our bounds, we additionally estimate a divergence surrogate between $\pi_T(\cdot \mid s)$ and $\pi_K(\cdot \mid s)$ on held-out states $s \sim d$ via sample-based distances (e.g. MMD or sliced Wasserstein between action samples under matched noise seeds), and plot empirical return gap versus the estimated mismatch.

Goal-conditioned navigation under distributional shift (AntMaze). We next recommend AntMaze because it is both diffusion-relevant (multimodal actions are common near bottlenecks) and shift-sensitive (coverage gaps in $\mathcal{D}$ are typical). We train a teacher $\pi_T(a \mid s)$ on AntMaze variants (e.g. umaze/medium/large, with diverse goal locations), and distill students π_K. Evaluation should include (i) the standard test goal distribution and (ii) deliberate shifts d that emphasize states weakly represented in $\mathcal{D}$: goals near narrow passages, starts near walls, or altered goal distributions (e.g. "corner-only" goals). We report success rate and path length, and we stratify performance by a dataset-coverage proxy (e.g. nearest-neighbor distance from evaluation states to $\mathcal{D}$). This stratification is a practical way to examine whether latent alignment reduces mismatch amplification as one moves away from dataset support.

Missing-data navigation (maze2d with controlled coverage holes). To isolate the role of representation preservation, we recommend a synthetic missing-data protocol in maze2d-style environments. Starting from a full dataset $\mathcal{D}$, we create $\mathcal{D}_{\text{miss}}$ by deleting transitions in selected spatial regions ("holes") or by removing action modes (e.g. deleting samples with actions in a cone), while keeping the evaluation environment unchanged. The target distribution d is then chosen to place substantial mass inside or near these deleted regions (for instance, by sampling start/goal pairs that force traversal near holes). This construction allows us to test whether a student trained on $\mu = d_{\mathcal{D}_{\text{miss}}}$ can still match the teacher when queried on states that are

OOD relative to the training support. We expect that student variants with a latent alignment penalty $\eta_z > 0$ and multi-noise-level matching exhibit smaller degradation than students trained only on terminal actions.

Controlled OOD contextual bandit. We recommend an explicit contextual bandit experiment (horizon 1) to validate the tight linear dependence on action-distribution mismatch in a setting where confounders from multi-step dynamics are absent. We choose a continuous-action bandit with bounded rewards $r(s, a) \in [-R_{\max}, R_{\max}]$ and construct contexts s such that the teacher $\pi_T(\cdot \mid s)$ is intentionally multimodal (e.g. a mixture of Gaussians with context-dependent weights). The student π_K is obtained by distillation under varying budgets K and/or controlled under-training to sweep a range of mismatches ε. We then evaluate, for each d (including an OOD context distribution obtained by shifting context covariates), the empirical return gap and a high-confidence estimate of $\mathbb{E}_{s\sim d}[\mathrm{TV}(\pi_T, \pi_K)]$ (approximated via discretization or by bounding TV through $\sqrt{\mathrm{KL}}$ with Monte Carlo estimates). The expected outcome is an approximately linear relationship, consistent with the bandit bound, and a clear demonstration that improving distribution matching on shifted d is the relevant quantity for preserving reward.

Hardware-realistic evaluation (real robot or high-fidelity simulation). To substantiate the latency motivation, we recommend at least one deployment-oriented setting: either a real robot (e.g. tabletop pushing or pick-and-place) using a fixed offline dataset $\mathcal{D}$, or a hardware-realistic simulator (e.g. with actuation delays, sensor noise, and torque limits). The central measurement is closed-loop success under a fixed control frequency constraint. We report (i) success rate as a function of the allowed per-action latency budget, (ii) measured latency distributions on the intended compute platform (GPU and/or embedded CPU), and (iii) the tradeoff curve between K and performance under the same wall-clock budget. The intent is to demonstrate that a student with small K can meet real-time constraints while remaining close to the teacher's behavior.

Latency benchmarks and profiling protocol. For each policy we measure end-to-end action-sampling time, separating sequential depth from constant-factor overhead. We recommend reporting (a) per-action mean/percentiles, (b) effective control rate (Hz) under a fixed batch size, and (c) the scaling with K and T at fixed architecture. The primary comparison is π_T versus π_K, but we also include a one-step student when available. To avoid misleading conclusions, we fix compilation settings and use the same trunk width where possible, reporting both wall-clock time and the count of sequential network evaluations.

Ablations. We recommend ablations that identify which ingredients most influence OOD matching: (i) remove latent alignment ($\eta_z = 0$); (ii) match only terminal actions (drop intermediate score/noise matching); (iii) train a small-T diffusion policy directly on $\mathcal{D}$ (no teacher) to test whether "short diffusion" is a substitute for distillation; (iv) distill with a frozen trunk versus a trainable trunk to assess representation transfer; (v) vary teacher-timestep subsampling q_T and student auxiliary evaluations q_K to quantify the training-cost–fidelity tradeoff. For each ablation we report both return metrics and distribution-matching surrogates on in-distribution and shifted d, since improvements that only appear in $J(\pi)$ may not reflect faithful teacher matching.

9 Discussion and Limitations

Our contribution is a distillation procedure that reduces the sequential depth of SRDP-style diffusion policies while preserving, to the extent possible, the teacher's state-conditioned action distribution. Theorems 1–3 clarify what such a guarantee can and cannot buy: if we can enforce small conditional mismatch on a target state distribution d, then the induced return gap is small; conversely, in the worst case the dependence on this mismatch is necessarily linear (and amplified by $(1-\gamma)^{-2}$ in the discounted setting). We therefore view the student as an *accelerator* of the teacher rather than an independent offline RL agent: it inherits both the strengths and the blind spots of π_T, and it cannot transcend the information-theoretic limitations of strictly offline learning under shift.

Calibration under far-OOD shift. A recurring practical question is whether the student remains *calibrated* when queried on states that are far from the offline support. Even if we drive the empirical distillation loss on training states μ to (near) zero, it does not follow that $\mathbb{E}_{s\sim d}[\mathrm{TV}(\pi_T, \pi_K)]$ is small for a shifted d. Our latent-alignment regularizer is motivated by the hope that, when d remains within a neighborhood of the representation manifold learned by the teacher, controlling $\|z_T - z_K\|$ stabilizes generalization of the action map g_K (cf. Theorem 4). However, under *far*-OOD shift, two failure modes remain: (i) the teacher representation $z_T(s)$ itself may be non-informative or unstable, and (ii) the student may match an *incorrectly calibrated* teacher distribution extremely well. In such regimes, the relevant question is not only whether $\pi_K \approx \pi_T$, but also whether π_T is trustworthy. Since our framework treats π_T as an oracle, any calibration pathologies (e.g. overconfident unimodal actions where the task is multimodal) will be faithfully replicated by a successful student.

A second calibration issue concerns the *diffusion-time* axis. The teacher defines a family of conditional distributions over noisy actions a_t for $t =$

$1, \ldots, T$. When $K \ll T$, the student is effectively approximating a coarse-grained reverse-time dynamics. Matching only terminal actions can yield a student that is correct marginally at $t = 0$ on μ but behaves unpredictably under perturbations of the sampling process (e.g. different noise seeds or different initializations). Intermediate score matching and multi-noise-level latent alignment partially address this, yet they do not constitute a certificate that the student sampler is stable as a Markov chain. We view this as an open theoretical gap: our bounds are stated in terms of the *final* conditional action distribution, whereas practical deployment often depends on the numerical and stochastic stability of the sampling procedure.

Teacher failure modes and compounding effects. The hardness statement in the global context emphasizes that, under support mismatch, even the teacher may be unidentifiable from $\mathcal{D}$ alone. Distillation cannot repair such failures: if π_T takes catastrophically wrong actions on a region that is unseen in $\mathcal{D}$, then a student that matches π_T there will also be wrong. More subtly, a student may *amplify* teacher errors if approximation artifacts are correlated across states. Theorem 2 measures mismatch under d_{π_T}, but the student induces its own occupancy d_{π_K}; if small local mismatches cause the trajectory distribution to shift into regions where the student is less accurate, then the realized return gap can exceed what would be predicted from a mismatch estimate computed on an exogenous held-out set. This is not a defect of the bound so much as a reminder that occupancy coupling is the core difficulty in sequential decision-making. In practice, this motivates either (i) evaluating mismatch under mixture occupancies (as noted after Theorem 2), or (ii) incorporating a conservative training distribution μ that overweights states likely to be visited by π_K (without online rollouts, this can only be approximated).

Interaction with Q-guidance and other test-time control. Diffusion policies are often combined with test-time guidance mechanisms, such as using an estimated action-value $Q(s, a)$ to bias the reverse diffusion trajectory toward high-value actions. Distillation interacts with such guidance in two distinct ways. First, if the deployed policy is the *guided* teacher, then the natural target of distillation is the guided conditional distribution $\pi_T^{\text{guide}}(\cdot \mid s)$ rather than the unguided π_T. This is conceptually straightforward but may be expensive: guidance typically requires additional Q-evaluations per denoising step, and thus changes the teacher's computational graph. Second, one may wish to distill an *unguided* student and then apply guidance at test time. This preserves flexibility but reintroduces sequential cost, and it may break the distributional matching guarantee since the guided distribution is not the one minimized during training. Moreover, guidance can magnify small modeling errors in score estimates: if the guidance term is large, then

the reverse dynamics may be dominated by Q-gradients in regions where the student (or the Q estimator) is inaccurate. We therefore regard the "distill-then-guide" pathway as requiring additional safeguards, such as explicit regularization of the guidance strength, or joint distillation of both the policy and the guiding signal.

Implications for safety. Our guarantees are *behavioral* (match the teacher) rather than *normative* (satisfy task constraints). Even perfect matching of π_T does not imply safety if π_T violates safety constraints in rare or OOD circumstances. More pointedly, a small expected TV under a distribution d does not control worst-case behavior over all states, nor does it provide reachability guarantees for unsafe sets. Thus, distillation should be viewed as orthogonal to safety mechanisms such as control barrier functions, verified safety filters, or constraint-aware planning. A pragmatic integration is to deploy π_K behind a certified filter and to treat the filter-induced intervention rate as an additional metric: a student that matches the teacher but triggers the filter more often may be operationally worse, even if $J(\pi_K) \approx J(\pi_T)$ in unconstrained rollouts.

Future work: vision, partial observability, and certified abstention. Three extensions appear technically immediate but conceptually nontrivial. First, vision-based policies introduce a high-dimensional observation map $o \mapsto s$ (explicit or implicit), and the relevant "OOD" notion becomes ambiguous: shift may occur in pixel space while remaining in-distribution in latent space, or vice versa. Here, representation preservation could be strengthened by matching teacher and student latents at multiple network depths, and by calibrating uncertainty with respect to nuisance factors (lighting, viewpoint). Second, in POMDPs, the diffusion policy is naturally conditioned on a belief or history embedding; distillation then becomes a two-timescale problem in which errors in the history encoder affect all subsequent action distributions. Extending Theorem 4 to recurrent embeddings (with stability constants over time) would be a principled starting point. Third, we are interested in *certified abstention*: the student should be able to refuse action (or defer to a slower teacher/controller) when it cannot guarantee small mismatch on the encountered state. One approach is to learn an upper bound on $\mathrm{TV}(\pi_T, \pi_K)$ from observable quantities (e.g. latent distances $\|z_T - z_K\|$, or density estimates in z_T-space), and then to couple this bound with Theorem 2 to obtain a conservative return-loss certificate. Developing such a certificate with statistical validity under distribution shift, without online interaction, remains open and would materially improve the safety story of fast distilled diffusion control.