Value-Rectified Distillation for Flow-based Offline Reinforcement Learning

Ke Jiang[1,2,3], **Wen Jiang**[1,2], **Yoshinobu Kawahara**[3,4], **Xiaoyang Tan**[1,2]
[1]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
[2]MIIT Key Laboratory of Pattern Analysis and Machine Intelligence
[3]Graduate School of Information Science and Technology, The University of Osaka
[4]Center for Advanced Intelligence Project, RIKEN
{ke_jiang, darren.jum, x.tan}@nuaa.edu.cn,
kawahara@ist.osaka-u.ac.jp

Abstract

We aim to address the problem of *Out-of-Distribution (OOD) Mode Averaging*, a critical problem that emerges when distilling Flow Matching (FM) based behavior policies in offline Reinforcement Learning (RL): When learning over multi-modal datasets, standard distillation inadvertently forces the policy to interpolate across different behavioral modes, causing the generated actions to collapse into invalid, OOD regions. To resolve this, we propose **Value-Rectified Distillation (VRD)**, a novel framework that reformulates offline policy distillation as a value-guided generative trajectory alignment problem. By employing a dynamic coupling mechanism, VRD actively rectifies flow dynamics to explicitly decouple distinct behavioral modes, cleanly isolating high-value actions. Specifically, we instantiate VRD through two complementary algorithms: **VRD-Q**, which utilizes value-quantile behavior distillation to statistically filter out low-value behaviors, and **VRD-R**, which reorganizes the latent space to geometrically isolate high-value generative trajectories from the FM-based behavior policy. Extensive empirical evaluations demonstrate that VRD improves existing generative baselines across diverse continuous control and planning tasks on the D4RL, OGBench, and Minari benchmarks, especially over those multi-modal, non-expert datasets.

1 Introduction

Offline Reinforcement Learning (RL) aims to derive a high-capacity policy from a fixed, pre-collected dataset without further environment interaction [16, 15], which is particularly vital for the applications where online exploration is costly or hazardous. Recently, generative models, particularly Flow Matching (FM) [18], have emerged as a powerful alternative to traditional Diffusion-based methods [9, 2, 5], where FM-based policies, such as Flow Q-Learning (FQL) [21], offer significant advantages in modeling complex, multi-modal behavior distributions while circumventing the training instabilities typically associated with backpropagating through ODE solvers.

Despite their empirical promise, we identify a fundamental pathology in how current flow-based RL frameworks distill high-value actions from multi-modal behavior policies. Specifically, when learning from multi-modal data, our analysis reveals that Flow Matching networks inherently suffer from *Latent Fusion*: the topological entanglement between the generative trajectories of optimal and sub-optimal behavioral modes, caused by the vector field's training dynamics, as is shown in Figure 1 (left). In previous works, like FQL [21], it first pre-trains a behavior flow policy to capture the full multi-modal behavior distribution [4] from the dataset, and subsequently applies Q-value guidance to distill the optimal policy. However, because such distillation operates on a continuous

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

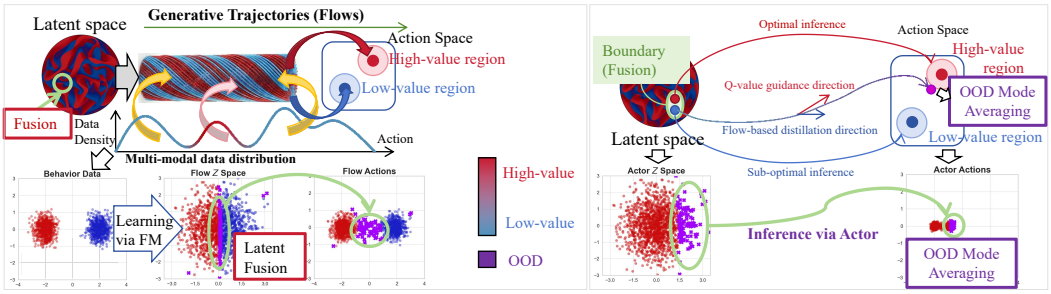


Figure 1: Problem of *Latent Fusion* (left) and the OOD Mode Averaging problem caused by FQL-style behavior distillation (right).

vector field where trajectories are already topologically entangled by the MLE pre-training, the Q-guided objective, as is shown in Figure 1 (right), inadvertently forces the sampled actions to linearly interpolate across distinct modes during inference, which triggers the *OOD Mode Averaging* disaster, where it is risky for the generated actions to collapse into low-density regions of the state-action manifold, ultimately causing performance degradation. Although recent approaches like Guided Flow Policy (GFP) [24] incorporate reward information to amplify optimal behavior modes in FM-based behavior policy, their FQL-style distillation mechanism still suffers from *OOD Mode Averaging*.

To address this issue, our core insight is that during the distillation phase, policy refinement must go beyond simply cloning from FM-based behavior policy; rather, it inherently requires rectifying the flow dynamics to explicitly **decouple distinct behavioral modes** along the data manifold. Therefore, we introduce **Value-Rectified Distillation (VRD)**, a novel framework that reformulates offline policy distillation as a value-guided trajectory alignment problem, of which the core lies in a dynamic coupling mechanism that precisely modulates the flow dynamics to explicitly decouple and exclude sub-optimal behavioral modes, ensuring that the generated actions are confined to the optimal action manifold during inference.

To instantiate this unified framework, we derive two distinct algorithms from complementary mathematical perspectives: **VRD-Q (Statistical Rectification)**: this variant performs in-place vector field alignment by employing asymmetric weight modulation based on value quantiles, which statistically filters sub-optimal regions without altering the underlying flow topology. **VRD-R (Geometric Rectification)**: this variant actively reshapes the flow's topology, which constructs a value-gravity field that transports the generative trajectories toward a highly compact optimal support set, geometrically isolating them from sub-optimal modes.

In summary, our primary contributions include: **1) Tracing OOD Mode Averaging:** We theoretically trace the root cause of the *OOD Mode Averaging* problem in FQL-style distillation mechanism, revealing how its distillation objective interacts with *latent fusion* to generate invalid OOD actions on multi-modal data. **2) The VRD Framework:** We propose Value-Rectified Distillation (VRD) with two implementations VRD-Q and VRD-R, reformulating offline policy distillation as a value-guided trajectory alignment problem to safely extract high-value policies without mode averaging. **3) Empirical Validation:** Extensive evaluations demonstrate that VRD-Q and VRD-R effectively mitigate this OOD issue, outperforming or performing competitively with existing generative baselines across various continuous control and planning tasks.

2 Preliminaries and motivation

2.1 Offline RL and Flow Q-Learning

Suppose a Markov Decision Process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, \rho, p)$, then the goal of offline RL is to learn a policy π that maximizes the expected cumulative discounted return using a static dataset $\mathcal{D}$ collected by a behavioral policy. Previous work adopted the FQL framework [21] to avoid the instability of backpropagating through ODE solvers. FQL consists of three components:

Critic: A value function $Q_\phi(s, a)$ trained to minimize the Bellman error.

BC Flow Policy: A flow-matching policy $\mu_\theta(s, z)$ defined by a vector field v_θ, trained purely on Behavioral Cloning (BC) to capture the dataset distribution:

$$\mathcal{L}_{\text{Flow}}(\theta) = \mathbb{E}_{s,a\sim\mathcal{D},t\sim\mathcal{U}(0,1),x^0\sim\mathcal{N}(0,I_d)}[\|v_\theta(t, s, x^t) - (a - x^0)\|_2^2] \tag{1}$$

where $x^t = (1 - t)x^0 + ta$.

One-Step Actor Policy: An expressive feed-forward network $\pi(s, z) : \mathcal{S} \times \mathbb{R}^d \to \mathcal{A}$ that maps latent noise z directly to actions.

In standard FQL, the actor π is trained to maximize Q-values while being regularized via distillation from the frozen flow policy μ_θ. The standard objective uses an unstructured Gaussian prior $z \sim \mathcal{N}(0, I_d)$ for this coupling:

$$\mathcal{L}_\pi = \mathbb{E}_{s\sim\mathcal{D},z\sim\mathcal{N}(0,I_d)}[-\lambda \cdot Q_\phi(s, \pi(s, z)) + \|\pi(s, z) - \mu_\theta(s, z)\|_2^2] \tag{2}$$

, while the Q-networks Q_ϕ are learned by minimizing a standard double Temporal Difference (TD) objective:

$$\mathcal{L}(\phi_i) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(Q_{\phi_i}(s, a) - r - \gamma \min_{j=1,2} Q_{\bar{\phi},j}(s', \pi(s', z))\right)^2\right],\ z \sim \mathcal{N}(0, I) \tag{3}$$

This implies that the latent space of the actor is isotropic; the position of z carries no semantic information regarding the quality of the action.

2.2 Latent fusion and OOD mode-averaging problems of FQL in multi-modal data

According to Flow Q-Learning, we can fit the action distribution of the behavior policy $\pi_\beta(a|s)$, by the flow model $\mu_\theta(s, z)$, where $z \sim \mathcal{N}(0, I)$. Then we define the support action set given s, as $\mathcal{A}(s) = \{a|\pi_\beta(a|s) > 0\}$, and OOD action set $\bar{\mathcal{A}} = \{a|\pi_\beta(a|s) = 0\}$. Furthermore, given a Q-function $Q(s, a)$ as the evaluator, we can define the optimal and suboptimal action sets as,

$$\mathcal{A}_Q^*(s) = \{a|a \in \mathcal{A}(s), Q(s, a) \geq \max_{a'} Q(s, a') - \epsilon_{opt}\} \tag{4}$$

$$\mathcal{A}_Q^-(s) = \mathcal{A}(s)/\mathcal{A}_Q^*(s) \tag{5}$$

In our work, we focus on the setting of **multi-modal data**, which has raised attention in the field [3, 10] recently. Specifically, in our work, we focus on the **isolated** multi-modal data as,

$$\inf_{a^*\in\mathcal{A}_Q^*(s),a^-\in\mathcal{A}_Q^-(s)} \|a^* - a^-\|_2 = D > 0 \tag{6}$$

which means that there exists a gap between the optimal and sub-optimal data modes. Under such kind of data, the first problem we encounter in flow-based offline RL is the *Latent Fusion* for the flow policy μ_θ learned by $\mathcal{L}_{\text{Flow}}(\theta)$ in Eq.(1):

Proposition 1 (Latent Entanglement of OOD Generation)**.** *Suppose the latent space $\mathcal{Z}$ (support of $z \sim \mathcal{N}(0, I)$) is connected, and the flow mapping $\mu_\theta(s, z)$ is L-Lipschitz continuous w.r.t. z. Let $\mathcal{Z}_Q^*(s)$, $\mathcal{Z}_Q^-(s)$, and $\mathcal{Z}_{OOD}(s)$ denote the latent preimages of $\mathcal{A}_Q^*(s)$, $\mathcal{A}_Q^-(s)$, and $\bar{\mathcal{A}}(s)$, respectively. If the model expresses both optimal and sub-optimal modes ($\mathcal{Z}_Q^*(s) \neq \emptyset, \mathcal{Z}_Q^-(s) \neq \emptyset$), the following holds: **(1)** The optimal latent boundary is non-empty ($\partial\mathcal{Z}_Q^*(s) \neq \emptyset$); **(2)** The OOD latent set is strictly non-empty ($\mathcal{Z}_{OOD}(s) \neq \emptyset$); **(3)** Any ϵ-neighborhood of a boundary point $z_{bd} \in \partial\mathcal{Z}_Q^*(s)$ intersects non-optimal regions: $\forall\epsilon>0, \mathbb{B}_\epsilon(z_{bd}) \cap \left(\mathcal{Z}_Q^-(s) \cup \mathcal{Z}_{OOD}(s)\right) \neq \emptyset$.*

Proof is shown in Appendix A.1. Proposition 1 formalizes the *Latent Fusion* dilemma. Because μ_θ is continuous and $\mathcal{Z}$ is connected, the model cannot safely isolate optimal latents. Property (2) proves that to bridge the physical gap $D > 0$, any continuous curve connecting the modes must structurally intersect the OOD space (a topological necessity, though without quantifying probability mass). Furthermore, Property (3) shows $\mathcal{Z}_Q^*(s)$ is inextricably entangled with sub-optimal/OOD regions at its boundary.

While Proposition 1 leverages the potent multi-modal modeling capabilities of Flow matching for targeted policy extraction, the extreme topological entanglement within the latent space, as characterized by the *Latent Fusion* in Proposition 1, renders such distillation and mapping via continuous, smooth networks susceptible to mode-averaging, as shown in Theorem 1. This phenomenon forces the policy to generate interpolated, OOD actions ($\bar{a} \notin \mathcal{A}(s)$) that collapse into off-manifold regions.

Theorem 1 (OOD Mode Averaging in FQL)**.** *Suppose the Q-function satisfies μ-strong concavity in the region bridging the modes and is L-Lipschitz smooth. Consider the FQL-style distillation that minimizes the regularized objective in Eq.(2). Under multi-modal data with gap $D > 0$ as Eq.(6) and any non-zero value regularization penalty $\lambda > 0$, FQL generates OOD actions with a minimum expected error bounded by $D^2/4$.*

Proof is shown in Appendix A.2. Theorem 1 reveals a fundamental limitation of FQL: under the latent mode entanglement established in Proposition 1, the explicit Euclidean regularization (e.g., $-\lambda Q + \|a - \mu_\theta\|_2^2$) blindly forces policy updates to perform linear interpolation across the action space. Because multi-modal data manifolds are topologically separated by a physical gap $D > 0$, this shortest Euclidean path traverses low-density OOD valleys, which triggers the Mode Averaging disaster with a strict expected error bounded by $D^2/4$. Consequently, we conclude that action-space Euclidean constraints are structurally incompatible with multi-modal offline RL. Besides, assuming local μ-strong concavity in the valley between modes serves as an analytical abstraction to capture the optimization tension in FQL.

In other words, in probability density estimation, a smoothly connected set of feasible actions (like slightly different stable walking postures) topologically constitutes a single, broad, and continuous mode. Flow Matching naturally and gracefully captures this continuous variance. However, we suggest that arises specifically when the behavioral distribution contains structurally distinct peaks separated by low-density valleys (OOD regions).

Motivation. The observation that standard Euclidean distillation is vulnerable to mode averaging on multi-modal datasets motivates our Value-Rectified Distillation (VRD) framework. As analyzed in this section, directly penalizing the action difference can inadvertently pull the actor into OOD regions. To effectively mitigate this issue, policy extraction must move beyond simple action-space regularization. Instead of linearly interpolating actions across empty spaces, VRD rectifies the continuous latent mapping by dynamically aligning the flow dynamics with value guidance, hence encouraging the policy to safely extract high-value actions along the data manifold.

3 Methodology

3.1 Unified framework: kernel-coupled value rectification

In Flow Q-Learning, while the initial behavior cloning phase effectively captures the multi-modal data distribution via the flow model μ_θ, it absorbs suboptimal behavioral modes (as discussed in Section 2). To distill μ_θ into an optimal policy π that exclusively outputs actions within the optimal set $\mathcal{A}_Q^*$, a principled rectification mechanism is essential. Rather than treating this process as isolated heuristic regularizations, we cast the value rectification as a latent distribution alignment problem.

To this end, we propose a generalized Value-Rectified Functional Space. We define the rectification objective $\Psi(\pi; \mu_\theta)$ as an expected kernel-coupled divergence over the joint latent space $\mathcal{Z} \times \mathcal{Z}$:

$$\Psi(\pi; \mu_\theta) = \int_{\mathcal{Z}} \mathcal{P}_{\mathcal{Z}}(z) \left[\int_{\mathcal{Z}} \kappa_Q(z, z' \mid s, \mu_\theta, \pi) \cdot \|\mu_\theta(s, z) - \pi(s, z')\|_2^2 \, dz' \right] dz \tag{7}$$

where $\mathcal{P}_{\mathcal{Z}}(z)$ is the prior distribution of the pre-trained latent space, and $\kappa_Q(z, z' \mid \cdot) : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}^+$ is defined as the Value-Guided Alignment Kernel (VGAK). Specifically, in our work, we focus on a *Non-merging transport kernel* $\kappa_Q(z, z' \mid \cdot) = w(z)\delta(z' - T(z))$ for a measurable weight $w \geq 0$ and a measurable map $T : \mathcal{Z} \to \mathcal{Z}$ satisfying mode separation $T(\mathcal{Z}_\mu^*) \cap T(\mathcal{Z}_\mu^-) = \emptyset$ and fiber-wise optimal preservation (i.e., for any $z' \in T(\mathcal{Z}_\mu^*)$, the induced conditional barycenter over the pre-image fiber $T^{-1}(\{z'\})$ remains within $\mathcal{A}_Q^*(s)$). The core insight of this unified functional is that the transition from a behavior-cloned flow to an optimal policy is essentially a latent-to-latent coupling process. The kernel κ_Q mathematically governs both the coupling strength (how much weight a sample carries) and the coordinate mapping (where the latent variable is transported) between the source latent variable z and the target latent variable z'.

Then, by plugging Eq. (7) into Eq. (2), the rectification-based policy optimization objective is:

$$\mathcal{L}_\pi = \mathbb{E}_{s \sim \mathcal{D}} \left[\Psi(\pi; \mu_\theta \mid s) - \lambda \cdot \mathbb{E}_{z' \sim \mathcal{P}_{test}}[Q(s, \pi(s, z'))] \right] \tag{8}$$

where λ is the balance coefficient, and $\mathcal{P}_{test}$ is a prior distribution of the latent variable z' used at test.

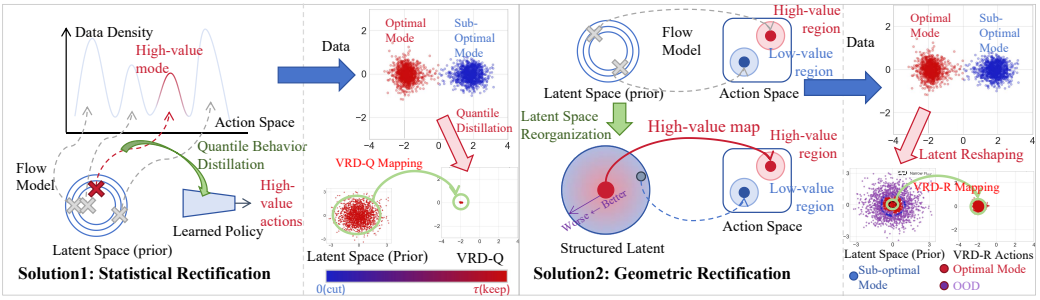


Figure 2: Implementations for **Value-Rectified Distillation (VRD)**: Statistical rectification via quantile behavior distillation (left), Geometric rectification via reshaping behavior distillation (right).

To establish a rigorous theoretical foundation for how this functional guarantees optimal policy extraction, we first define the partitioning of the original latent space.

Definition 1 (Optimal Latent Subset)**.** *For a given state s and a pre-trained flow model μ_θ, the subset of latent variables that map to the optimal action set $\mathcal{A}_Q^*(s)$ in Eq. (4) is defined as $\mathcal{Z}_\mu^*(s) = \{z \in \mathcal{Z} \mid \mu_\theta(s,z) \in \mathcal{A}_Q^*(s)\}$, and the corresponding suboptimal latent subset is $\mathcal{Z}_\mu^-(s) = \mathcal{Z} \setminus \mathcal{Z}_\mu^*(s)$.*

Theorem 2 (Optimal Support)**.** *Suppose the critic Q is conservative on out-of-distribution regions such that $\sup_{a \in \overline{\mathcal{A}}(s)} Q(s,a) < \max_{a' \in \mathcal{A}(s)} Q(s,a') - \epsilon_{opt}$. Let $\hat{\pi}^* = \arg\min_\pi [\Psi(\pi;\mu_\theta) - \lambda \cdot \mathbb{E}_{z' \sim \mathcal{P}_{test}}[Q(s,\pi(s,z'))]]$, as in Eq. (8), be the learned optimal policy assuming universal approximation. Over the test support $\mathcal{Z}_{test} = \mathrm{supp}(\mathcal{P}_{test})$, if the non-merging alignment kernel κ_Q satisfies: (i) **ϵ-Bounded Suboptimal Leakage:** $\int_{\mathcal{Z}_{test}} \int_{\mathcal{Z}_\mu^-} \mathcal{P}_\mathcal{Z}(z)\kappa_Q(z,z' \mid \cdot)\,dzdz' \le \epsilon$, and (ii) **$C$-Dominant Optimal Coupling:** $\int_{\mathcal{Z}_{test}} \int_{\mathcal{Z}_\mu^*} \mathcal{P}_\mathcal{Z}(z)\kappa_Q(z,z' \mid \cdot)\,dzdz' \ge C > 0$, then under the kernel-induced marginal distribution over $\mathcal{Z}_{test}$, the policy guarantees to output optimal actions with probability $\mathbb{P}(\hat{\pi}^*(s,z') \in \mathcal{A}_Q^*(s)) \ge \frac{C}{C+\epsilon}$.*

Proof see Appendix A.3. Theorem 2 resolves this issue by entirely discarding the Euclidean penalty in the action space. Instead, it reformulates the policy distillation process as a non-linear latent distribution alignment problem, introducing a Value-Guided Alignment Kernel, denoted as κ_Q. As long as κ_Q satisfies two fundamental conditions: (i) ϵ-Bounded Suboptimal Leakage, and (ii) C-Dominant Optimal Coupling, Theorem 2 guarantees a positive lower bound ($\ge \frac{C}{C+\epsilon}$) on the probability that the distilled policy outputs actions within the optimal support set $\mathcal{A}_Q^*(s)$. Mathematically, this non-linear latent alignment allows the generative trajectories to conform to the underlying data manifold, safely bypassing OOD regions and eliminating the mode averaging.

In the following subsections (Section 3.2 and 3.3), we demonstrate how the single framework, VRD, derives two practical rectification methodologies, as shown in Figure 2, one via quantile modulation and the other via latent reshaping, by simply instantiating different forms of Dirac kernels κ_Q.

3.2 Statistical rectification: quantile behavior distillation

In our first instantiation of the generalized framework, we restrict the target latent space from undergoing topological deformation. From an Eulerian perspective, we perform an **in-place alignment** where the source latent z must strictly align with the target latent z'.

Mathematically, this implies that the Value-Guided Alignment Kernel (VGAK) degenerates into a diagonal Dirac matrix modulated by a value-aware weight. We formulate this specific kernel as:

$$\kappa_Q(z,z' \mid s,\mu_\theta,\pi) = \underbrace{\left|\tau - \mathbb{I}\big(Q(s,\mu_\theta(s,z)) < Q(s,\pi(s,z)) - \epsilon_{\text{opt}}\big)\right|}_{\text{Quantile-aware Weight } \omega_\tau(z)} \cdot \underbrace{\delta(z'-z)}_{\text{In-place Alignment}} \tag{9}$$

where $\tau \in (0.5, 1)$ is a quantile hyperparameter, $\mathbb{I}(\cdot)$ is the indicator function, and $\delta(\cdot)$ is the Dirac delta function. Note that the weight $\omega_\tau(z)$ is strictly a function of the source latent z, ensuring stable

gradient propagation during variational optimization. In practice, ϵ_{opt} is often set to a very small value, e.g., 10^{-6}.

By substituting this diagonal kernel into the unified functional $\Psi(\pi; \mu_\theta)$ (defined in Section 3.1), the inner integral over z' collapses due to the sifting property of the Dirac δ function. The functional seamlessly derives into our first practical objective, the **Quantile-Rectified Loss**:

$$\Psi_Q(\pi; \mu_\theta) = \mathbb{E}_{s\sim\mathcal{D}, z\sim\mathcal{Z}}\left[\left|\tau - \mathbb{I}\Big(Q\big(s, \mu_\theta(s,z)\big) < Q\big(s, \pi(s,z)\big) - \epsilon_{\text{opt}}\Big)\right| \cdot \Big\|\mu_\theta(s,z) - \pi(s,z)\Big\|_2^2\right] \tag{10}$$

Proposition 2 (Connecting VRD-Q with Theorem 2)**.** *Let P^* and P^- be the prior masses of $\mathcal{Z}_\mu^*$ and $\mathcal{Z}_\mu^-$, respectively. For given constants $\epsilon > 0$ and $C > 0$, if the quantile threshold $\tau \in (0.5, 1)$ satisfies $\tau P^* \geq C$ and $(1-\tau)P^- + (2\tau - 1)P^-_{band}(\tau) \leq \epsilon$ (where $P^-_{band}(\tau)$ denotes the prior mass of the boundary band near $\mathcal{A}_Q^*$), then: (a) The VRD-Q kernel satisfies the coupling conditions of Theorem 2, guaranteeing $\mathbb{P}_{\kappa_Q}(\hat{\pi}^*(s,z) \in \mathcal{A}_Q^*(s)) \geq \frac{C}{C+\epsilon}$ under the kernel-induced marginal distribution; (b) Under the standard prior $z \sim \mathcal{N}(0, I_d)$ at inference, $\mathbb{P}_z(\hat{\pi}^*(s,z) \in \mathcal{A}_Q^*(s)) \geq P^* + P^-_{rect}(\tau)$, where $P^-_{rect}(\tau)$ is the rectified prior mass.*

Proof see Appendix A.4. Proposition 2 has two parts. Part (a) analyzes the kernel-induced marginal distribution: by the margin in Eq. (9), optimal latents receive weight τ independent of π, while suboptimal ones outside a band near $\mathcal{A}_Q^*(s)$ receive $1-\tau$. Hence, Theorem 2's coupling conditions hold if $\tau P^* \geq C$ and $(1-\tau)P^- + (2\tau-1)P^-_{\text{band}}(\tau) \leq \epsilon$. Part (b) covers inference under the standard prior without latent reweighting: optimal latents still map to $\mathcal{A}_Q^*(s)$, and a weakened behavioral anchor rectifies some suboptimal latents into $\mathcal{A}_Q^*(s)$, yielding the lower bound $P^* + P^-_{\text{rect}}(\tau)$. Notably, this bound is distinct and does not follow from (a)'s $\frac{C}{C+\epsilon}$ bound.

Practically, tuning τ trades off suboptimal filtering (larger τ heavily attenuates suboptimal weights) against critic-error robustness (moderate $1-\tau$ anchors the policy against noisy Q-gradients); see Appendix C.3 for empirical τ analysis.

3.3 Geometric rectification: reshaped behavior distillation

In our second instantiation, instead of modulating the fitting weights, we assign equal importance to all samples but permit topological deformation of the target latent space. This achieves rectification through **spatial transportation**, where the source latent z is dynamically routed to a new coordinate z' based on its evaluated Q-value.

Mathematically, the Value-Guided Alignment Kernel (VGAK) degenerates into a non-diagonal, deterministic transport map. We formulate this transport kernel as a shifted Dirac delta function:

$$\kappa_Q(z, z' \mid s, \mu_\theta, \pi) = 1 \cdot \delta(z' - z_Q(s, z, \mu_\theta)) \tag{11}$$

where $z_Q(s, z, \mu_\theta)$ is a value-guided reshaping function. To construct an optimal cluster that is highly accessible during inference, we design z_Q as a gravitational field towards high-value actions:

$$z_Q(s, z, \mu_\theta) = \Phi_Q(s, z) \cdot \frac{z}{\|z\|_2}, \quad \text{where} \quad \Phi_Q(s, z) = \frac{Q_{\max}(s) - Q(s, \mu_\theta(s, z))}{Q_{\max}(s) - Q_{\min}(s) + \epsilon_m} \tag{12}$$

where $Q_{\max}(s) = \max_{a\in\mathcal{A}(s)} Q(s,a)$ and $Q_{\min}(s) = \min_{a\in\mathcal{A}(s)} Q(s,a)$ denote the state-wise value boundaries (which can be practically approximated by the empirical extremes or the MDP reward bounds $\frac{R_{\max}}{1-\gamma}$ and $\frac{R_{\min}}{1-\gamma}$). Here, $\Phi_Q(s,z) \in [0,1]$ acts as a normalized ranking scalar. By substituting this transport kernel into the unified functional $\Psi(\pi; \mu_\theta)$ in Eq. (7), the integral domain collapses precisely at $z' = z_Q(s, z, \mu_\theta)$. This derivation seamlessly yields our second practical objective, the Reshaping-Rectified Loss:

$$\Psi_R(\pi; \mu_\theta) = \mathbb{E}_{s\sim\mathcal{D}, z\sim\mathcal{Z}}\left[\|\mu_\theta(s,z) - \pi(s, z_Q(s,z,\mu_\theta))\|_2^2\right] \tag{13}$$

Proposition 3 (Connecting VRD-R with Theorem 2)**.** *Suppose the critic Q satisfies the conservatism condition in Theorem 2, and the optimal action set $\mathcal{A}_Q^*(s)$ forms a convex dominant mode over which $Q(s,\cdot)$ is concave. During inference, let $z' \sim \mathcal{N}(0, \sigma^2 I)$ be truncated to $\mathcal{Z}_{test} = \{z' \in \mathcal{Z} \mid$*

$\|z'\|_2 \leq \sigma\}$. *If the truncation radius satisfies* $\sigma < \epsilon_{opt}/(Q_{\max}(s) - Q_{\min}(s) + \epsilon_m)$, *where* ϵ_{opt} *is the optimality relaxation factor in Eq. (4), then* κ_Q *in Eq. (11) fulfills all conditions of Theorem 2 with zero suboptimal leakage* ($I^- = 0$), *guaranteeing* $\mathbb{P}(\hat{\pi}^*(s, z') \in \mathcal{A}_Q^*(s)) = 1$ *under the kernel-induced marginal distribution.*

Proof see Appendix A.5. Unlike VRD-Q, whose threshold τ must be carefully balanced against probability masses to satisfy the ϵ and C bounds (Proposition 2), VRD-R translates the distribution alignment problem into a strict geometric isolation. By setting the inference truncation radius $\sigma < \epsilon_{\text{opt}}/(Q_{\max}(s) - Q_{\min}(s) + \epsilon_m)$, under the assumption of concave dominant optimal mode, VRD-R achieves the ideal extreme of Theorem 2 where the suboptimal leakage is approximate zero ($\epsilon \approx 0$). Consequently, the topological boundary $\epsilon_{\text{opt}}/(Q_{\max}(s) - Q_{\min}(s) + \epsilon_m)$ is determined exclusively by the Q-value landscape.

Proposition 3 bridges the geometric mapping of VRD-R with the requirements of Theorem 2, providing the precise analytical threshold for the truncation radius σ necessary to eliminate suboptimal leakage. By satisfying these conditions, it ensures that the inference process is not merely a heuristic reshaping but a value-guided optimal sampler that yields actions within $\mathcal{A}_Q^*(s)$ with high probability.

In summary, both VRD-Q and VRD-R prevent catastrophic out-of-distribution (OOD) mode-averaging by using dynamic Q-guidance to distill optimal actions along the pre-trained behavior manifold. However, they differ in mechanism and trade-offs: **VRD-Q (In-place Dynamical Intervention):** Truncates ODE gradient flows toward sub-optimal modes in-place via quantile modulation. It preserves the spatial geometry and multi-modal diversity of expert behaviors, which makes VRD-Q ideal for complex, multi-modal tasks requiring diverse path preservation and trajectory stitching (e.g., OGBench navigation). **VRD-R (Geometric Reshaping):** Rescales high-value latent vectors into a compact region around the origin. It acts as a symmetry-breaker to isolate a single, highly stable peak mode, which enable VRD-R for continuous control tasks where single-mode precision and stability are prioritized (e.g., standard D4RL locomotion). *Algorithm procedures of* ***VRD-Q*** *and* ***VRD-R*** *are illustrated in Appendix B. Hyperparameter analysis for both components see Appendix C.3.*

4 Related works

Gaussian Policies. Traditional offline RL methods often rely on Gaussian distributions to parameterize the policy. This category includes standard Behavior Cloning (**BC** [17]) as well as prominent value-based algorithms like Implicit Q-Learning (**IQL** [14]) and **ReBRAC** [23]. While Gaussian policies are computationally efficient, in complex offline datasets where behavior is often multi-modal or skewed, Gaussian-based actors suffer from limited expressivity [21], often leading to mode-averaging and suboptimal performance when the data distribution is non-Gaussian.

Diffusion Policies. To overcome the expressivity bottleneck, recent works such as **IDQL** [9], **SRPO** [2], and **CAC** [5] leverage the power of score-based diffusion modeling to capture highly complex and multi-modal action distributions. However, directly optimizing an iterative diffusion process to maximize a Q-function remains challenging, often requiring unstable backpropagation through the sampling steps or complex reward-guided sampling techniques [12].

Flow Policies. Flow matching [18] has emerged as a compelling alternative to diffusion, which is simpler to train and faster to infer. To address the difficulty of directly optimizing iterative flow models, Flow Q-Learning (**FQL**) [21] proposed distilling an expressive one-step policy from a flow-matching model guided by RL objectives, along with two variants: **FAWAC** [1] and **FBRAC** [25], which bridges the gap between expressivity and efficiency. More recently, Guided Flow Policy (**GFP**) [24] and Decision Flow (**DF**) [11] develops tailored optimization techniques, focusing on specifically extracting high-value actions within the flow framework.

5 Experiments

We aim to systematically assess the efficacy of the proposed VRD[1] by answering the core research questions (RQs): **RQ1 (Toy Example):** Can VRD-Q statistically filter sub-optimal modes and

[1]Our code is available at: `https://github.com/Jack10843/VRD-master`

VRD-R geometrically isolate optimal trajectories? **RQ2 (Benchmarks):** Does the overall VRD framework outperform existing baselines across standard continuous control (D4RL) and complex long-horizon planning (OGBench) tasks? **RQ3 (Multi-Modal):** How effectively does VRD mitigate OOD Mode Averaging when explicitly evaluated on different scales of multi-modal datasets? **RQ4 (Hyperparameters):** How robust is the VRD to key hyperparameter variations?

5.1 Visualizing latent rectification on toy examples (RQ1)

To empirically evaluate how our methods manipulate the latent space to extract optimal policies, we construct a 2D bi-modal toy environment. The dataset comprises actions sampled from an optimal mode at $[-2.0, 0.0]$ and a sub-optimal mode at $[2.0, 0.0]$, with isotropic Gaussian noise, where the corresponding value landscape is simulated using a mock Q-function strictly centered at $[-2.0, 0.0]$. Although the baseline flow model perfectly replicates the offline dataset by indiscriminately absorbing

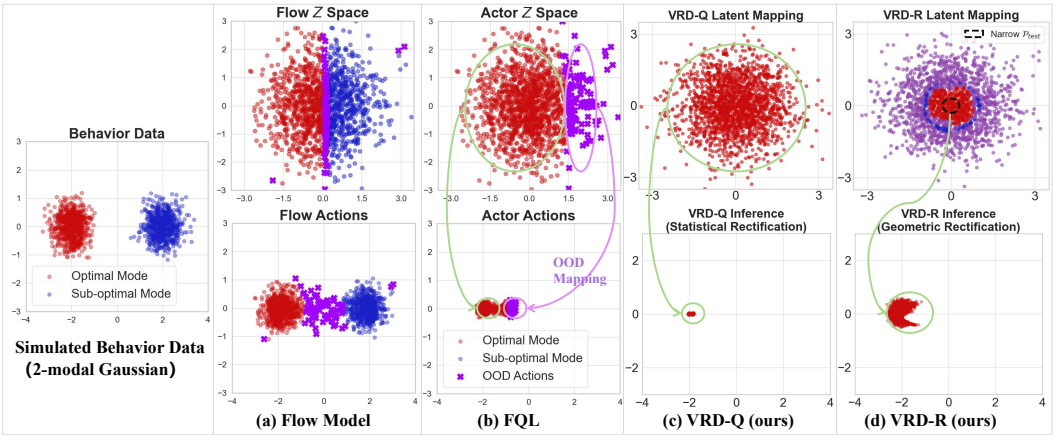


Figure 3: Visualizing results on toy examples with 2D multi-modal data.

both modes (Figure 3a), it suffers from *latent fusion* at the boundary between the optimal (red) and sub-optimal (blue) latent variables (Z), where the entanglement induces OOD action interpolations, hindering clear optimal mode distillation and exacerbating the risk of OOD mode averaging in FQL-style behavior distillation (Figure 3b). In contrast, the proposed VRD successfully extract a strictly optimal policy (Figure 3c, d). Specifically, VRD-Q performs statistical filtering by dynamically truncating distillation weights in low-value latent regions, which severs the gradient flow for sub-optimal actions, forcing the standard Gaussian prior to map exclusively to the optimal mode. Conversely, VRD-R geometrically reshapes the latent topology via value-guided spatial transportation. It attracts optimal latent variables toward the origin while expelling sub-optimal ones to the periphery. By sampling from a low-temperature Gaussian prior during zero-shot inference, VRD-R isolates optimal trajectories, verifying the efficacy of both rectification strategies.

5.2 Continuous control (D4RL) and complex planning (OGBench) tasks (RQ2)

D4RL [7] features **MuJoCo** locomotion (HalfCheetah, Hopper, Walker2d) across 'medium' (collected by a partially trained SAC [8]), 'medium-replay' (training history of SAC), and 'medium-expert' (mixed of expert data and 'medium') levels. It also includes **AntMaze** navigation across umaze, medium, and large layouts, using 'play' (undirected) and 'diverse' (goal-reaching) data to test trajectory stitching; and the 24-DOF **Adroit** Pen task with 'human' (demos), 'expert' (RL-trained), and 'cloned' (BC-mixed) datasets to evaluate high-dimensional control under narrow distributions.

As shown in Table 1, VRD-R outperforms the competitive baselines over most benchmarks, achieving the highest average score of **92.8**, while VRD-Q achieves **88.8**. On MuJoCo's hopper-med-expert (multi-modal), VRD-R scores **113.5** compared to FQL's 105.4, suggesting geometric reshaping effectively isolates expert modes to mitigate mode averaging. For sparse-reward navigation in antmaze-large-diverse, VRD-Q reaches **85.7** where diffusion policy methods like CAC fail (0.0), implying statistical filtering successfully truncates low-value actions for robust trajectory stitching.

Table 1: Results on D4RL MuJoCo, AntMaze and Adroit Benchmarks. The highest results of each row are bolded. Each result of our method is calculated over 8 seeds. Part of the results were reported in [21].

	Gaussian Policies			Diffusion Policies			Flow Policies						
Task	BC	IQL	ReBRAC	IDQL	SRPO	CAC	FAWAC	FBRAC	FQL	GFP	DF	**VRD-Q**	**VRD-R**
halfcheetah-med-expert	55.2	86.7	**101.1**	95.9	92.2	84.3	44.5	54.4	86.1	92.3	87.6	93.7 ± 1.2	**101.3** ± 1.3
hopper-med-expert	52.5	91.5	107.0	108.6	100.1	100.4	88.4	76.5	105.4	110.1	112.6	96.4 ± 13.5	**113.5** ± 0.4
walker2d-med-expert	107.5	109.6	111.6	112.7	114.0	110.4	57.3	53.5	102.7	106.4	111.3	112.1 ± 2.3	**113.7** ± 0.2
halfcheetah-medium	42.6	47.4	**65.6**	51.0	60.4	69.1	48.3	53.1	60.0	62.3	49.3	**65.7** ± 1.2	61.7 ± 4.5
hopper-medium	52.9	66.3	**102.0**	65.4	95.5	80.7	81.2	52.1	67.3	84.7	97.1	65.8 ± 15.8	76.3 ± 12.1
walker2d-medium	75.3	78.3	82.5	82.5	84.4	83.1	42.1	81.4	77.2	83.1	87.2	82.3 ± 4.9	**89.7** ± 6.4
halfcheetah-med-replay	36.6	44.2	51.0	45.9	51.4	**58.7**	42.6	49.6	53.1	54.7	44.2	54.7 ± 2.7	56.6 ± 2.1
hopper-med-replay	18.1	94.7	98.1	92.1	101.2	99.7	73.4	50.8	78.4	82.1	**102.3**	71.7± 15.1	96.3 ± 4.7
walker2d-med-replay	26.0	73.9	77.3	85.1	84.6	79.5	37.5	32.2	72.9	80.1	86.6	84.4 ± 7.1	**87.6** ± 5.3
Average (MuJoCo)	51.9	77.0	**88.5**	82.1	87.1	85.1	57.3	56.0	78.1	84.0	86.5	80.8	**88.5**
antmaze-umaze	55.0	77.0	98.0	94.0	97.0	66.0	90.0	94.0	96.0	96.8	93.7	96.7 ± 0.3	**99.3** ± 0.7
antmaze-umaze-diverse	47.0	62.2	84.0	80.0	82.0	66.0	55.0	62.0	89.0	91.9	83.2	83.2 ± 3.1	**94.2** ± 2.1
antmaze-medium-play	0.0	71.2	**90.0**	84.0	81.0	49.0	52.0	77.0	78.0	81.9	43.8	76.7 ± 9.5	87.5 ± 4.1
antmaze-medium-diverse	1.0	70.0	84.0	85.0	75.0	0.0	44.0	60.0	71.0	61.6	40.1	**87.1** ± 3.1	83.2 ± 0.7
antmaze-large-play	0.0	39.6	52.0	64.0	54.0	0.0	10.0	32.0	84.0	82.6	47.3	**84.4** ± 1.8	78.1 ± 8.3
antmaze-large-diverse	0.0	47.5	64.0	68.0	54.0	0.0	16.0	20.0	83.0	84.1	52.1	**85.7** ± 2.4	83.7 ± 5.7
Average (AntMaze)	17.2	61.3	78.7	79.2	73.8	30.2	44.5	57.5	83.5	83.2	60.0	85.6	**87.7**
pen-human	71	78	103	76	69	64	67	77	53	64.6	102.8	110.4± 2.7	**111.3** ± 4.3
pen-cloned	52	83	103	64	61	56	62	67	74	77.1	104.3	**108.1** ± 3.6	106.7 ± 6.7
pen-expert	110	128	**152**	140	134	103	118	119	142	140.4	139.1	138.4 ± 7.3	129.4 ± 12.9
Average (Adroit)	77.7	96.3	**119.3**	93.3	88.0	74.3	82.3	87.7	89.7	94.0	115.4	119.0	115.8
Average (ALL)	44.6	75.0	90.3	83.0	82.8	65.0	57.2	61.8	81.8	85.4	82.5	88.8	**92.8**

Finally, VRD-R achieves **111.3** on Adroit's pen-human (vs. GFP's 64.6). Overall, value-rectification allows generative policies to stably extract high-value manifolds even from mixed data distributions.

OGBench: To further evaluate our method's scalability to long-horizon and high-dimensional goal-conditioned tasks, we benchmark VRD-Q on 50 distinct tasks from **OGBench** [20]. These environments span exceptionally challenging navigation scenarios, including large and giant **AntMazes**, medium and large **HumanoidMazes**, and **AntSoccer**, as well as complex manipulation and spatial reasoning tasks like **Cube** (single/double), **Scene**, and sliding **Puzzles** (3x3/4x4).

Table 2: Results on OGBench Tasks. The highest results of each row are bolded. Each result of our method is calculated over 8 seeds. The full results are shown in Table 4 in Appendix, where part of the results were reported in [21].

	Gaussian Policies			Diffusion Policies			Flow Policies						
Task	BC	IQL	ReBRAC	IDQL	SRPO	CAC	FAWAC	FBRAC	IFQL	FQL	GFP	**VRD-Q**	**VRD-R**
antmaze-large	11.0	53.0	81.0	21.0	11.0	33.0	6.0	60.0	28.0	79.0	93.8	**95.5** ± 1.2	**94.9** ± 1.9
antmaze-giant	0.0	4.0	26.0	0.0	0.0	0.0	0.0	4.0	3.0	9.0	27.9	**49.8**± 5.3	38.3 ± 7.2
humanoidmaze-medium	2.0	33.0	22.0	1.0	1.0	53.0	19.0	38.0	60.0	58.0	72.0	**75.3**± 2.9	70.3± 7.2
humanoidmaze-large	1.0	2.0	2.0	1.0	0.0	0.0	0.0	2.0	11.0	4.0	17.8	**19.7** ± 2.9	16.8 ±5.9
antsoccer-arena	1.0	8.0	0.0	12.0	1.0	2.0	12.0	16.0	33.0	60.0	57.9	**63.8**± 4.7	47.9 ± 9.7
cube-single	5.0	83.0	91.0	95.0	80.0	85.0	81.0	79.0	79.0	96.0	**98.8**	**99.5** ± 0.4	97.7 ± 0.8
cube-double	2.0	7.0	12.0	15.0	2.0	6.0	5.0	15.0	14.0	29.0	**47.2**	45.6 ± 6.4	32.2 ± 8.4
scene	5.0	28.0	41.0	46.0	20.0	40.0	30.0	45.0	30.0	**56.0**	53.5	**55.4** ± 3.2	**55.8** ± 2.5
puzzle-3x3	2.0	9.0	21.0	10.0	18.0	19.0	6.0	14.0	19.0	30.0	23.1	28.6 ± 3.8	**34.7** ± 11.6
puzzle-4x4	0.0	7.0	14.0	29.0	10.0	15.0	1.0	13.0	25.0	17.0	26.1	**30.6** ± 3.5	19.8 ± 6.3
Average (50 tasks)	2.8	23.4	31.0	22.8	14.3	25.3	16.1	28.6	30.0	43.5	51.8	**56.4**	50.8

As shown in Table 2, VRD-Q achieves the highest overall average score of **56.4** across all 50 tasks, demonstrating superior robustness over existing baselines. Notably, in antmaze-giant, VRD-Q (49.8) and VRD-R (38.3) outperform the best Flow baseline, GFP (27.9), while Diffusion-based policies, like CAC, fail (0.0). VRD-Q also performs well over high-dimensional locomotion (humanoidmaze, antsoccer-arena), whereas VRD-R (50.8 average) excels in spatial reasoning tasks like puzzle-3x3 (34.7) and scene (55.8). Ultimately, both VRD-Q's statistical filtering and VRD-R's geometric reshaping effectively mitigate compounding OOD errors, scaling robustly to large-scale challenges where standard unconditional policy distillation methods collapse.

5.3 Deep dive into multi-modal datasets (RQ3)

To evaluate robustness against multi-modal distributions, we construct mixed datasets combining expert (optimal) and random (sub-optimal) trajectories, as discussed in [13, 27]. Maintaining a 1,000,000-sample buffer, we vary the ratio of random samples from 50% to 90%, progressively

diminishing the expert mode's presence, which serves as a rigorous stress test to determine if our mechanisms can successfully isolate the vanishing optimal mode from multi-modal data.

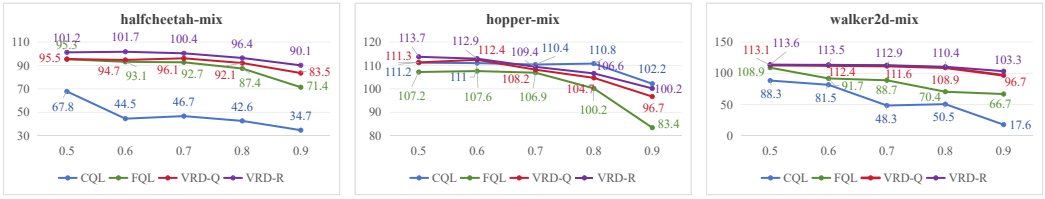


Figure 4: Results over different scales of multi-modal datasets.

From Figure 4, we observe that both the statistical filtering of VRD-Q and the geometric reshaping of VRD-R exhibit a more gradual performance degradation as the proportion of sub-optimal modes (i.e., the random sample ratio) increases. Specifically, on the 'halfcheetah-mix' and 'walker2d-mix' datasets, both VRD variants significantly outperform the CQL and FQL baselines. This improved robustness is likely because agents in these two environments are more susceptible to overfitting to sub-optimal noisy data, a vulnerability that the proposed VRD methods appear to effectively mitigate.

5.4 Hyperparameter analysis (RQ4)

We evaluate the sensitivity of the quantile parameter τ in VRD-Q and the scale parameter σ of latent variables in VRD-R's inference. From Table 3, VRD-Q performance follows an inverted U-shape, peaking at $\tau = 0.7$ to balance noise exclusion with distribution coverage. VRD-R achieves its optimum at $\sigma = 10^{-4}$, where moderate transformation effectively decouples modes without distorting latent topology. We recommend $\tau = 0.7$ and $\sigma = 10^{-4}$ as default settings. Specifically, τ should be based on sub-optimal data proportions, while σ should be tailored to the specific characteristics of environments.

Table 3: Sensitivity analysis for main hyperparameters τ of VRD-Q and σ of VRD-R ($\lambda = 0.05$, full results see Appendix C.3).

	VRD-Q (τ)					**VRD-R** (σ)				
Values	0.5	0.6	0.7	0.8	0.9	1	1e-2	1e-3	1e-4	1e-5
halfcheetah-m	60.4	63.4	**65.7**	65.3	58.1	45.7	60.6	60.4	**61.7**	61.2
hopper-m	64.3	65.1	**65.8**	65.1	56.4	50.2	73.9	75.5	**76.3**	74.2
walker2d-m	76.5	81.7	**82.3**	81.1	77.4	53.9	89.1	88.6	**89.7**	88.3

Other experimental details see Appendix C, e.g., Minari results (Appendix C.2), full hyperparameter analysis (Appendix C.3) and validation study for core components of VRD (Appendix C.4).

6 Conclusion

This paper proposes the Value-Rectified Distillation (VRD) framework, reformulating policy distillation as a value-guided trajectory alignment process to solve the *latent fusion* problem in offline RL over multi-modal dataset. VRD-Q and VRD-R mitigate OOD Mode Averaging problem via statistical filtering and geometric reshaping respectively. Experiments show VRD successfully decouples the high-value behavior mode from the multi-modal data, which enables VRD outperform baselines on various benchmarks and datasets, demonstrating exceptional robustness for sub-optimal data noise.

Limitations (see Appendix D). The proposed VRD relies on the underlying value network's accuracy, where severe Q-value overestimation may mislead the alignment process. Future work could explore more robust value estimation or extend this framework to discrete action spaces.

Acknowledgements

This work was partially supported by National Natural Science Foundation of China (62476128). This work was partially supported by JSPS KAKENHI Grant Number JP26H02498, and JST Moonshot R&D Program JPMJMS263K.

References

[1] Nair Ashvin, Dalal Murtaza, Gupta Abhishek, and L Sergey. Accelerating online reinforcement learning with offline datasets. *CoRR, vol. abs/2006.09359*, 2020.

[2] Huayu Chen, Cheng Lu, Zhengyi Wang, Hang Su, and Jun Zhu. Score regularized policy optimization through diffusion behavior. *arXiv preprint arXiv:2310.07297*, 2023.

[3] Tianyu Chen, Zhendong Wang, and Mingyuan Zhou. Diffusion policies creating a trust region for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 37:50098–50125, 2024.

[4] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

[5] Zihan Ding and Chi Jin. Consistency models as a rich and efficient policy class for reinforcement learning. *arXiv preprint arXiv:2309.16984*, 2023.

[6] Gerald B Folland. *Real analysis: modern techniques and their applications*. John Wiley & Sons, 1999.

[7] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

[8] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

[9] Philippe Hansen-Estruch, Ilya Kostrikov, Michael Janner, Jakub Grudzien Kuba, and Sergey Levine. Idql: Implicit q-learning as an actor-critic method with diffusion policies. *arXiv preprint arXiv:2304.10573*, 2023.

[10] Zhang-Wei Hong, Aviral Kumar, Sathwik Karnik, Abhishek Bhandwaldar, Akash Srivastava, Joni Pajarinen, Romain Laroche, Abhishek Gupta, and Pulkit Agrawal. Beyond uniform sampling: Offline reinforcement learning with imbalanced datasets. *Advances in neural information processing systems*, 36:4985–5009, 2023.

[11] Jifeng Hu, Sili Huang, Siyuan Guo, Zhaogeng Liu, Li Shen, Lichao Sun, Hechang Chen, Yi Chang, and Dacheng Tao. Decision flow policy optimization. *arXiv preprint arXiv:2505.20350*, 2025.

[12] Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022.

[13] Ke Jiang, Jia-Yu Yao, and Xiaoyang Tan. Recovering from out-of-sample states via inverse dynamics in offline reinforcement learning. *Advances in Neural Information Processing Systems*, 36:38815–38826, 2023.

[14] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

[15] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in neural information processing systems*, 32, 2019.

[16] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33:1179–1191, 2020.

[17] Aviral Kumar, Joey Hong, Anikait Singh, and Sergey Levine. Should i run offline reinforcement learning or behavioral cloning? In *International conference on learning representations*, 2021.

[18] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[19] James Munkres. Topology james munkres second edition, 2014.

[20] Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. Ogbench: Benchmarking offline goal-conditioned rl. *arXiv preprint arXiv:2410.20092*, 2024.

[21] Seohong Park, Qiyang Li, and Sergey Levine. Flow q-learning. *arXiv preprint arXiv:2502.02538*, 2025.

[22] Walter Rudin. Principles of mathematical analysis. 2021.

[23] Denis Tarasov, Vladislav Kurenkov, Alexander Nikulin, and Sergey Kolesnikov. Revisiting the minimalist approach to offline reinforcement learning. *Advances in Neural Information Processing Systems*, 36:11592–11620, 2023.

[24] Franki Nguimatsia Tiofack, Théotime Le Hellard, Fabian Schramm, Nicolas Perrin-Gilbert, and Justin Carpentier. Guided flow policy: Learning from high-value actions in offline reinforcement learning. *arXiv preprint arXiv:2512.03973*, 2025.

[25] Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

[26] Omar G. Younis, Rodrigo Perez-Vicente, John U. Balis, Will Dudley, Alex Davey, and Jordan K Terry. Minari, September 2024.

[27] Hongchang Zhang, Jianzhun Shao, Yuhang Jiang, Shuncheng He, Guanwen Zhang, and Xiangyang Ji. State deviation correction for offline reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 9022–9030, 2022.

A Proofs

A.1 Proof of Proposition 1

Proposition 1. *Suppose the latent space $\mathcal{Z}$ (support of $z \sim \mathcal{N}(0, I)$) is connected, and the flow mapping $\mu_\theta(s, z)$ is L-Lipschitz continuous w.r.t. z. Let $\mathcal{Z}_Q^*(s)$, $\mathcal{Z}_Q^-(s)$, and $\mathcal{Z}_{OOD}(s)$ denote the latent preimages of $\mathcal{A}_Q^*(s)$, $\mathcal{A}_Q^-(s)$, and $\bar{\mathcal{A}}(s)$, respectively. If the model expresses both optimal and sub-optimal modes ($\mathcal{Z}_Q^*(s) \neq \emptyset, \mathcal{Z}_Q^-(s) \neq \emptyset$), the following holds: **(1)** The optimal latent boundary is non-empty ($\partial\mathcal{Z}_Q^*(s) \neq \emptyset$); **(2)** The OOD latent set is strictly non-empty ($\mathcal{Z}_{OOD}(s) \neq \emptyset$); **(3)** Any ϵ-neighborhood of a boundary point $z_{bd} \in \partial\mathcal{Z}_Q^*(s)$ intersects non-optimal regions: $\forall \epsilon > 0, \mathbb{B}_\epsilon(z_{bd}) \cap \left(\mathcal{Z}_Q^-(s) \cup \mathcal{Z}_{OOD}(s)\right) \neq \emptyset$.*

Proof. Due to the prior latent space $\mathcal{Z} = \mathbb{R}^d$ (often a standard Gaussian $\mathcal{N}(0, I)$) is path-connected, there exists a continuous path $\gamma : [0, 1] \to \mathcal{Z}$ connecting the pre-images:

$$\gamma(0) = z^* \in \mu_\theta^{-1}(\mathcal{A}_Q^*), \quad \gamma(1) = z^- \in \mu_\theta^{-1}(\mathcal{A}_Q^-) \tag{14}$$

Let the mapped trajectory be $\Lambda = \gamma([0, 1])$. Because Λ is a connected set that intersects both the interior of $\mathcal{Z}_Q^*(s)$ and its exterior complement, it cannot be partitioned into two disjoint open sets. By the fundamental property of connected spaces [19] (Section 23, Exercise 6), Λ must intersect the topological boundary of $\mathcal{Z}_Q^*(s)$. Therefore, there exists at least one boundary point $z_{bd} \in \partial\mathcal{Z}_Q^*(s)$ along the path, strictly proving **claim (1)**: $\partial\mathcal{Z}_Q^*(s) \neq \emptyset$.

By composition of continuous functions $\mu_\theta(s, z)$, the mapped trajectory Γ in the action space is a connected set [22] (Theorem 4.22):

$$\Gamma = \{\mu_\theta(s, \gamma(t)) \mid t \in [0, 1]\} \tag{15}$$

Since Γ is connected and connects $\mathcal{A}_Q^*$ and $\mathcal{A}_Q^-$ across a geometric gap D, Γ cannot be entirely contained within the disconnected support $\mathcal{A} = \mathcal{A}_Q^* \cup \mathcal{A}_Q^-$ [19] (Lemma 23.2). Define the OOD latent set:

$$\mathcal{Z}_{OOD} = \{z \in \mathcal{Z} \mid \mu_\theta(s, z) \in \bar{\mathcal{A}}\} \tag{16}$$

By the Intermediate Value Theorem [22] (Theorem 4.23) on connected spaces, $\Gamma \cap \bar{\mathcal{A}} \neq \emptyset \implies \mathcal{Z}_{OOD} \neq \emptyset$, hence proving **claim (2)**.

Then, observe that the action space is partitioned into three mutually disjoint sets: the optimal mode $\mathcal{A}_Q^*(s)$, the sub-optimal mode $\mathcal{A}_Q^-(s)$, and the OOD region $\bar{\mathcal{A}}(s)$. Because the flow mapping $\mu_\theta(s, \cdot)$ is a well-defined function assigning every latent z to exactly one action, its pre-images preserve this disjointness and form a mutually exhaustive partition of the entire latent space $\mathcal{Z}$:

$$\mathcal{Z} = \mathcal{Z}_Q^*(s) \cup \mathcal{Z}_Q^-(s) \cup \mathcal{Z}_{OOD}(s) \tag{17}$$

Consequently, the topological complement of the optimal pre-image in $\mathcal{Z}$ is exactly the union of the remaining regions: $\mathcal{Z} \setminus \mathcal{Z}_Q^*(s) = \mathcal{Z}_Q^-(s) \cup \mathcal{Z}_{OOD}(s)$.

According to the standard neighborhood characterization of a topological boundary [19] (Section 17, Exercise 19(b)), a point z_{bd} belongs to the boundary $\partial\mathcal{Z}_Q^*(s)$ if and only if every open set containing z_{bd} intersects both the set itself and its complement. Therefore, for any arbitrary radius $\epsilon > 0$, the open ϵ-ball $\mathbb{B}_\epsilon(z_{bd})$ must strictly intersect the complement $\mathcal{Z} \setminus \mathcal{Z}_Q^*(s)$ [19] (Theorem 17.5 (a)). Substituting the exact complement derived in Eq.(17) directly proves **claim (3)**:

$$\forall \epsilon > 0, \quad \mathbb{B}_\epsilon(z_{bd}) \cap \left(\mathcal{Z}_Q^-(s) \cup \mathcal{Z}_{OOD}(s)\right) \neq \emptyset \tag{18}$$

Completing the proof.

□

A.2 Proof of Theorem 1

Theorem 1. *Suppose the Q-function satisfies μ-strong concavity in the region bridging the modes and is L-Lipschitz smooth. Consider the FQL policy update $\pi(s, z)$ which minimizes the regularized objective in Eq.(2). Under multi-modal data with gap $D > 0$ as Eq.(6), FQL fails to recover the optimal mode and instead generates OOD actions with a minimum expected error bounded by $\frac{D^2}{4}$.*

Proof. We reformulate the objective in Eq.(2) as:

$$\mathcal{L}_\pi = \mathbb{E}_{s\sim\mathcal{D}, z\sim\mathcal{N}(0,I_d)}[-Q_\phi(s, \pi(s, z)) + \alpha \cdot \|\pi(s, z) - \mu_\theta(s, z)\|_2^2] \tag{19}$$

which switches the coefficient λ over value to α onto the behavior regularization. Based on the latent entanglement established in Proposition 1, consider a latent variable $z \in \mathcal{Z}_Q^-(s)$ whose pre-trained behavior action falls in the sub-optimal mode, i.e., $a_{bc}^- = \mu_\theta(s, z) \in \mathcal{A}_Q^-(s)$. The objective of FQL given state s aims to solve:

$$\min_a \left(-Q(s, a) + \alpha\|a - \mu_\theta(s, z)\|_2^2\right) \tag{20}$$

$$\Leftrightarrow \max_a \left(Q(s, a) - \alpha\|a - \mu_\theta(s, z)\|_2^2\right) \tag{21}$$

Denote the objective function as $\mathcal{J}(a) = Q(s, a) - \alpha\|a - a_{bc}^-\|_2^2$. Assuming Q is twice differentiable, the optimal action generated by FQL, denoted as a_{FQL}, must satisfy the first-order optimality condition $\nabla_a\mathcal{J}(a_{FQL}) = 0$:

$$\nabla_a Q(s, a_{FQL}) - 2\alpha(a_{FQL} - a_{bc}^-) = 0 \tag{22}$$

Let $a^* \in \mathcal{A}_Q^*(s)$ be the optimal action that maximizes Q, meaning $\nabla_a Q(s, a^*) = 0$. By the Mean Value Theorem (or assuming locally quadratic Q), we can express the gradient using the Hessian matrix $H = \nabla_a^2 Q(s, a)$:

$$\nabla_a Q(s, a_{FQL}) = \nabla_a Q(s, a_{FQL}) - \nabla_a Q(s, a^*) = H(a_{FQL} - a^*) \tag{23}$$

Plugging Eq.(23) into Eq.(22), we obtain:

$$H(a_{FQL} - a^*) - 2\alpha(a_{FQL} - a_{bc}^-) = 0 \tag{24}$$

Assuming the Q-function satisfies μ-strong concavity in the region bridging the modes and is L-Lipschitz smooth, there exists a symmetric positive-definite matrix $P = -H$ such that $\mu I \preceq P \preceq LI$. Substituting P:

$$\begin{aligned} & -P(a_{FQL} - a^*) - 2\alpha(a_{FQL} - a_{bc}^-) = \mathbf{0} \\ \Rightarrow \quad & (2\alpha I + P)a_{FQL} = Pa^* + 2\alpha a_{bc}^- \end{aligned} \tag{25}$$

Because $P \succ 0$ and $2\alpha > 0$, the matrix $(2\alpha I + P)$ is strictly positive-definite and thus invertible. Therefore, we can explicitly solve for a_{FQL}:

$$a_{FQL} = (2\alpha I + P)^{-1}Pa^* + (2\alpha I + P)^{-1}(2\alpha I)a_{bc}^- \tag{26}$$

Let the weight matrix be $M(\alpha) = (2\alpha I+P)^{-1}P$. Recognizing that $M(\alpha)+(2\alpha I+P)^{-1}(2\alpha I) = I$, the FQL action is precisely a matrix-weighted convex combination (interpolation) of the two modes:

$$a_{FQL} = M(\alpha)a^* + (I - M(\alpha))a_{bc}^- \tag{27}$$

Consider the trajectory of a_{FQL} as the Euclidean regularization coefficient $\alpha \in (0, +\infty)$ varies, forming a continuous parametric curve $\gamma(\alpha) \subset \mathbb{R}^{\dim(\mathcal{A})}$:

- When $\alpha \to 0$ (no behavior regularization), $M(\alpha) \to I$, yielding $\lim_{\alpha\to 0}\gamma(\alpha) = a^* \in \mathcal{A}_Q^*(s)$.
- When $\alpha \to +\infty$ (strict behavior cloning), $M(\alpha) \to \mathbf{0}$, yielding $\lim_{\alpha\to\infty}\gamma(\alpha) = a_{bc}^- \in \mathcal{A}_Q^-(s)$.

As established in Proposition 1, the multi-modal valid support sets $\mathcal{A}_Q^*(s)$ and $\mathcal{A}_Q^-(s)$ are strictly separated by an Euclidean distance $D > 0$. By the connectedness theorem in topology, the continuous spatial curve $\gamma(\alpha)$ connecting these two separated modes must inherently traverse the low-density OOD vacuum region $\bar{\mathcal{A}}(s) = \{a \mid \pi_\beta(a|s) = 0\}$.

Specifically, there exists a regularization strength α where the interpolation is geometrically centered between the modes (e.g., when the eigenvalues of $M(\alpha)$ are roughly 0.5). At this intersection, the generated action a_{FQL} is positioned in the middle of the physical gap D. Consequently, its minimum Euclidean distance to *any* valid action $a \in \mathcal{A}(s)$ is bounded below by $\frac{D}{2}$:

$$\min_{a \in \mathcal{A}(s)} \|a_{FQL} - a\|_2 \geq \frac{\|a^* - a_{bc}^-\|_2}{2} \geq \frac{D}{2} \tag{28}$$

This strictly implies that the expected minimum squared prediction error incurred by the FQL mode averaging is bounded by:

$$\min_{a \in \mathcal{A}(s)} \|a_{FQL} - a\|_2^2 \geq \left(\frac{D}{2}\right)^2 = \frac{D^2}{4} > 0 \tag{29}$$

This proves that the action-space Euclidean penalty structurally forces FQL to output invalid OOD actions with a strict error lower bound, completing the proof. □

A.3 Proof of Theorem 2

Theorem 2. Suppose the critic Q is conservative on out-of-distribution regions such that $\sup_{a \in \overline{\mathcal{A}}(s)} Q(s,a) < \max_{a' \in \mathcal{A}(s)} Q(s,a') - \epsilon_{opt}$. Let $\hat{\pi}^* = \arg\min_\pi [\Psi(\pi; \mu_\theta) - \lambda \cdot \mathbb{E}_{z' \sim \mathcal{P}_{test}}[Q(s, \pi(s,z'))]]$, as in Eq. (8), be the learned optimal policy assuming universal approximation. Over the test support $\mathcal{Z}_{test} = \mathrm{supp}(\mathcal{P}_{test})$, if the non-merging alignment kernel κ_Q satisfies (i) ϵ-Bounded Suboptimal Leakage: $\int_{\mathcal{Z}_{test}} \int_{\mathcal{Z}_\mu^-} \mathcal{P}_\mathcal{Z}(z) \kappa_Q(z, z' \mid \cdot)\, dz dz' \leq \epsilon$; and (ii) C-Dominant Optimal Coupling: $\int_{\mathcal{Z}_{test}} \int_{\mathcal{Z}_\mu^*} \mathcal{P}_\mathcal{Z}(z) \kappa_Q(z, z' \mid \cdot)\, dz dz' \geq C > 0$, then under the kernel-induced marginal distribution over $\mathcal{Z}_{\text{test}}$, the policy guarantees to output optimal actions with probability $\mathbb{P}(\hat{\pi}^*(s,z') \in \mathcal{A}_Q^*(s)) \geq \frac{C}{C+\epsilon}$.

Proof. To analyze the exact properties of the learned optimal policy $\hat{\pi}^*$ without relying on global unimodal convexity assumptions, we perform variational calculus on the regularized global objective in the function space. The proof proceeds in the following rigorous mathematical steps:

First, we expand the objective functional in Eq. (8) into integral forms over the latent priors $\mathcal{P}_\mathcal{Z}(z)$ and $\mathcal{P}_{test}(z')$. By applying Fubini-Tonelli's Theorem ([6], Theorem 2.37) to swap the integration order, we isolate the inner pointwise objective for any given $z' \in \mathrm{supp}(\mathcal{P}_{test})$:

$$\mathcal{L}(\pi(s,z')) = \int_\mathcal{Z} \mathcal{P}_\mathcal{Z}(z) \kappa_Q(z, z' \mid \cdot) \|\mu_\theta(s,z) - \pi(s,z')\|_2^2\, dz - \lambda \mathcal{P}_{test}(z') Q(s, \pi(s,z')) \tag{30}$$

Let $Z_{norm}(z') = \int_\mathcal{Z} \mathcal{P}_\mathcal{Z}(z) \kappa_Q(z, z' \mid \cdot)\, dz$ denote the marginal coupling density at z', and let $\mathcal{W}(z \mid z') = \mathcal{P}_\mathcal{Z}(z) \kappa_Q(z, z' \mid \cdot) / Z_{norm}(z')$ be the induced conditional probability distribution over the source latent space $\mathcal{Z}$. Treating the alignment kernel κ_Q as a fixed guidance operator during the variational evaluation step (consistent with the stop-gradient formulation in Section 3.1), dividing Eq. (30) by $Z_{norm}(z') > 0$ shows that the optimal action $\hat{\pi}^*(s,z')$ maximizes the conditionally weighted functional:

$$\hat{\pi}^*(s,z') = \arg\max_a \left\{ Q(s,a) - \frac{1}{2\alpha(z')} \int_\mathcal{Z} \mathcal{W}(z \mid z') \|\mu_\theta(s,z) - a\|_2^2\, dz \right\}, \quad \alpha(z') = \frac{\lambda \mathcal{P}_{test}(z')}{2 Z_{norm}(z')} \tag{31}$$

Then we evaluate the optimality of $\hat{\pi}^*(s,z')$ strictly within the local fiber supported by $\mathcal{W}(\cdot \mid z')$. Specifically, for any candidate action a_{cand} constructed from the support of $\mathcal{W}(\cdot \mid z')$, comparing the objective value in Eq. (31) at the global maximizer $a = \hat{\pi}^*(s,z')$ with its value at $a = a_{cand}$ yields:

$$\begin{aligned} Q(s, \hat{\pi}^*(s,z')) - \frac{1}{2\alpha(z')} \int_\mathcal{Z} \mathcal{W}(\tilde{z} \mid z') \|\mu_\theta(s,\tilde{z}) - \hat{\pi}^*(s,z')\|_2^2\, d\tilde{z} \\ \geq Q(s, a_{cand}) - \frac{1}{2\alpha(z')} \int_\mathcal{Z} \mathcal{W}(\tilde{z} \mid z') \|\mu_\theta(s,\tilde{z}) - a_{cand}\|_2^2\, d\tilde{z} \end{aligned} \tag{32}$$

Next, consider drawing (z, z') from the effective joint coupling distribution over $\mathcal{Z} \times \text{supp}(\mathcal{P}_{test})$ with density proportional to $\mathcal{P}_{\mathcal{Z}}(z)\kappa_Q(z, z' \mid \cdot)$. Let $I^* = \int_{\text{supp}(\mathcal{P}_{test})} \int_{\mathcal{Z}_\mu^*} \mathcal{P}_{\mathcal{Z}}(z)\kappa_Q(z, z' \mid \cdot)\, dz dz'$ and $I^- = \int_{\text{supp}(\mathcal{P}_{test})} \int_{\mathcal{Z}_\mu^-} \mathcal{P}_{\mathcal{Z}}(z)\kappa_Q(z, z' \mid \cdot)\, dz dz'$ denote the effective optimal and suboptimal coupling masses over the test support, respectively. By conditions (i)–(ii) ($I^- \leq \epsilon$ and $I^* \geq C > 0$) and the monotonicity of $g(x, y) = \frac{x}{x+y}$:

$$\mathbb{P}(z \in \mathcal{Z}_\mu^*) = \frac{I^*}{I^* + I^-} \geq \frac{C}{C + \epsilon} \tag{33}$$

It remains to show that $z \in \mathcal{Z}_\mu^*$ implies $\hat{\pi}^*(s, z') \in \mathcal{A}_Q^*(s)$. Recall from Section 3.1 that the alignment kernel is formulated as a non-merging transport kernel $\kappa_Q(z, z' \mid \cdot) = w(z)\delta(z' - T(z))$ with disjoint target images $T(\mathcal{Z}_\mu^*) \cap T(\mathcal{Z}_\mu^-) = \emptyset$. For any fixed target latent $z' \in \text{supp}(\mathcal{P}_{test})$, the sifting property of the Dirac delta distribution restricts the effective support of $\kappa_Q(\cdot, z' \mid \cdot)$ over the source space $\mathcal{Z}$ strictly to the pre-image fiber:

$$\text{supp}_z(\kappa_Q(\cdot, z' \mid \cdot)) \subseteq T^{-1}(\{z'\}) = \{z \in \mathcal{Z} \mid T(z) = z'\} \tag{34}$$

Since $T(\mathcal{Z}_\mu^*) \cap T(\mathcal{Z}_\mu^-) = \emptyset$, the pre-image fiber $T^{-1}(\{z'\})$ in Eq. (34) can intersect at most one of the two subsets:

- If $z' \in T(\mathcal{Z}_\mu^*)$, then $z' \notin T(\mathcal{Z}_\mu^-) \Rightarrow T^{-1}(\{z'\}) \cap \mathcal{Z}_\mu^- = \emptyset$, which implies that $\kappa_Q(\cdot, z' \mid \cdot)$ vanishes identically on $\mathcal{Z}_\mu^-$.
- If $z' \in T(\mathcal{Z}_\mu^-)$, then $z' \notin T(\mathcal{Z}_\mu^*) \Rightarrow T^{-1}(\{z'\}) \cap \mathcal{Z}_\mu^* = \emptyset$, which implies that $\kappa_Q(\cdot, z' \mid \cdot)$ vanishes identically on $\mathcal{Z}_\mu^*$.
- If $z' \notin T(\mathcal{Z}_\mu^*) \cup T(\mathcal{Z}_\mu^-)$, $\kappa_Q(\cdot, z' \mid \cdot)$ vanishes on both subsets.

Consequently, for every $z' \in \text{supp}(\mathcal{P}_{test})$, $\kappa_Q(\cdot, z' \mid \cdot)$ vanishes on $\mathcal{Z}_\mu^-$ or on $\mathcal{Z}_\mu^*$, preventing any simultaneous non-zero coupling between optimal and suboptimal modes. Thus, up to a null event, $z \in \mathcal{Z}_\mu^*$ occurs if and only if $z' \in T(\mathcal{Z}_\mu^*)$, meaning $\text{supp}(\mathcal{W}(\cdot \mid z')) \subseteq \mathcal{Z}_\mu^*$.

Crucially, for any $z' \in T(\mathcal{Z}_\mu^*)$, every source latent $\tilde{z} \in \text{supp}(\mathcal{W}(\cdot \mid z'))$ belongs to $\mathcal{Z}_\mu^*$ and satisfies $\mu_\theta(s, \tilde{z}) \in \mathcal{A}_Q^*(s)$, i.e., $Q(s, \mu_\theta(s, \tilde{z})) \geq \max_{a'} Q(s, a') - \epsilon_{opt}$. Let $\bar{\mu}(s, z') = \int_{\mathcal{Z}} \mathcal{W}(\tilde{z} \mid z')\mu_\theta(s, \tilde{z})\, d\tilde{z}$ denote the conditional optimal barycenter over the fiber $T^{-1}(\{z'\})$. By applying the bias-variance decomposition:

$$\int_{\mathcal{Z}} \mathcal{W}(\tilde{z} \mid z')\|\mu_\theta(s, \tilde{z}) - a\|_2^2\, d\tilde{z} = \|a - \bar{\mu}(s, z')\|_2^2 + \text{Var}_{\mathcal{W}}(\mu_\theta) \tag{35}$$

to Eq. (32) and evaluating the objective against the candidate action $a_{cand} = \bar{\mu}(s, z')$, the constant variance term $\text{Var}_{\mathcal{W}}(\mu_\theta)$ cancels on both sides. We analyze the two resulting alignment regimes:

1. **Bijective / In-place Alignment (as in VRD-Q, $T(z) = z$):** The pre-image fiber in Eq. (34) is a singleton $T^{-1}(\{z'\}) = \{z'\}$, so $\mathcal{W}(\tilde{z} \mid z') = \delta(\tilde{z} - z')$ is a degenerate point mass at a single optimal action $\bar{\mu}(s, z') = \mu_\theta(s, z') \in \mathcal{A}_Q^*(s)$ with $\text{Var}_{\mathcal{W}}(\mu_\theta) = 0$. This preserves the exact multi-modal geometry of $\mathcal{A}_Q^*(s)$ point-by-point without any inter-mode averaging or convexity assumptions, reducing Eq. (32) directly to:

$$\begin{aligned} Q(s, \hat{\pi}^*(s, z')) &\geq Q(s, \mu_\theta(s, z')) + \frac{1}{2\alpha(z')}\|\hat{\pi}^*(s, z') - \mu_\theta(s, z')\|_2^2 \\ &\geq Q(s, \mu_\theta(s, z')) \geq \max_{a'} Q(s, a') - \epsilon_{opt} \end{aligned} \tag{36}$$

2. **Contractive / Reshaping Alignment (as in VRD-R):** When T compresses multiple optimal latents into the same target coordinate z', by the fiber-wise optimal preservation property of the non-merging kernel κ_Q (which we explicitly verify for VRD-R under a locally concave

optimal mode in Proposition 3), the conditional barycenter satisfies $\bar{\mu}(s, z') \in \mathcal{A}_Q^*(s)$. Substituting Eq. (35) with $a_{cand} = \bar{\mu}(s, z')$ into Eq. (32) yields:

$$\begin{aligned} Q(s, \hat{\pi}^*(s, z')) &\geq Q(s, \bar{\mu}(s, z')) + \frac{1}{2\alpha(z')} \|\hat{\pi}^*(s, z') - \bar{\mu}(s, z')\|_2^2 \\ &\geq Q(s, \bar{\mu}(s, z')) \geq \max_{a'} Q(s, a') - \epsilon_{opt} \end{aligned} \tag{37}$$

In either regime, combining Eq. (36) and Eq. (37), the event $z \in \mathcal{Z}_\mu^*$ (equivalently $z' \in T(\mathcal{Z}_\mu^*)$) directly guarantees $Q(s, \hat{\pi}^*(s, z')) \geq \max_{a' \in \mathcal{A}(s)} Q(s, a') - \epsilon_{opt}$. Furthermore, by the conservatism of the critic Q on out-of-distribution regions ($\sup_{a \in \overline{\mathcal{A}}(s)} Q(s, a) < \max_{a' \in \mathcal{A}(s)} Q(s, a') - \epsilon_{opt}$), this value lower bound strictly precludes $\hat{\pi}^*(s, z')$ from falling into $\overline{\mathcal{A}}(s)$, thereby guaranteeing $\hat{\pi}^*(s, z') \in \mathcal{A}(s)$ and hence $\hat{\pi}^*(s, z') \in \mathcal{A}_Q^*(s)$. Combined with Eq. (33), we obtain:

$$\mathbb{P}(\hat{\pi}^*(s, z') \in \mathcal{A}_Q^*(s)) \geq \mathbb{P}(z \in \mathcal{Z}_\mu^*) \geq \frac{C}{C + \epsilon} \tag{38}$$

Completing the proof. □

A.4 Proof of Proposition 2

Proposition 2. *Let P^* and P^- be the prior masses of $\mathcal{Z}_\mu^*$ and $\mathcal{Z}_\mu^-$, respectively. For given constants $\epsilon > 0$ and $C > 0$, if the quantile threshold $\tau \in (0.5, 1)$ satisfies $\tau P^* \geq C$ and $(1 - \tau)P^- + (2\tau - 1)P_{band}^-(\tau) \leq \epsilon$ (where $P_{band}^-(\tau)$ denotes the prior mass of the boundary band near $\mathcal{A}_Q^*$): (a) The VRD-Q kernel satisfies the coupling conditions of Theorem 2, guaranteeing $\mathbb{P}_{\kappa_Q}(\hat{\pi}^*(s, z) \in \mathcal{A}_Q^*(s)) \geq \frac{C}{C+\epsilon}$ under the kernel-induced marginal distribution; (b) Under the standard prior $z \sim \mathcal{N}(0, I_d)$ at inference, $\mathbb{P}_z(\hat{\pi}^*(s, z) \in \mathcal{A}_Q^*(s)) \geq P^* + P_{rect}^-(\tau)$, where $P_{rect}^-(\tau)$ is the rectified prior mass.*

Proof. Recall the VRD-Q kernel with optimality margin:

$$\kappa_Q(z, z' \mid s, \mu_\theta, \pi) = \underbrace{\left|\tau - \mathbb{I}\big(Q(s, \mu_\theta(s, z)) < Q(s, \pi(s, z)) - \epsilon_{\text{opt}}\big)\right|}_{\omega_\tau(z)} \cdot \delta(z' - z). \tag{39}$$

Since the weight depends on π, we use the stop-gradient fixed point of Theorem 2: $\hat{\pi}^*$ is a variational optimum if, for every z, $\hat{\pi}^*(s, z)$ maximizes the pointwise objective of Eq. (31) with $\omega_\tau(z)$ evaluated at $\pi = \hat{\pi}^*$. Because $T(z) = z$ and $\mathcal{P}_{\text{test}}(z) = \mathcal{P}_{\mathcal{Z}}(z)$, the penalty coefficient $\frac{1}{2\alpha(z)}$ in Eq. (31) equals $\alpha_{\text{reg}}(z) := \omega_\tau(z)/\lambda$, reducing Eq. (31) to

$$\hat{\pi}^*(s, z) = \arg\max_a \Big[Q(s, a) - \alpha_{\text{reg}}(z) \, \|a - \mu_\theta(s, z)\|_2^2\Big], \qquad \alpha_{\text{reg}}(z) = \omega_\tau(z)/\lambda. \tag{40}$$

Step 1: Closed form and improvement under the quadratic model. Write $v = \mu_\theta(s, z) - a^*$. For any regularization strength $\alpha > 0$, the first-order condition of Eq. (40), $-P(a - a^*) - 2\alpha(a - \mu_\theta(s, z)) = 0$, gives (cf. Eq. (26))

$$\hat{\pi}^*(s, z) - a^* = (I - M(\alpha))\, v, \qquad I - M(\alpha) = 2\alpha(2\alpha I + P)^{-1}. \tag{41}$$

Let $p_i \in [\mu_Q, L]$ be the eigenvalues of P, v_i the coordinates of v in its eigenbasis, and $r_i(\alpha) = \frac{2\alpha}{2\alpha + p_i} \in (0, 1)$. The value improvement of the actor over the behavior action is

$$\Delta(\alpha) := Q(s, \hat{\pi}^*(s, z)) - Q(s, \mu_\theta(s, z)) = \tfrac{1}{2} \sum_i p_i \big(1 - r_i(\alpha)^2\big) v_i^2. \tag{42}$$

Each $r_i(\alpha)$ increases with α, so $\Delta(\alpha)$ is strictly decreasing in α whenever $v \neq 0$. Moreover, $r_i(\alpha) \leq \frac{2\alpha}{2\alpha + \mu_Q}$, hence $1 - r_i(\alpha)^2 \geq c(\alpha)$ and

$$\Delta(\alpha) \geq c(\alpha) \cdot \tfrac{1}{2} v^\top P v = c(\alpha)\big(Q_{\max}(s) - Q(s, \mu_\theta(s, z))\big). \tag{43}$$

Note that the indicator in Eq. (39) equals 1 if and only if $\Delta(\alpha_{\text{reg}}(z)) > \epsilon_{\text{opt}}$.

Step 2: Weights on $\mathcal{Z}_\mu^*$. By critic conservatism, $\sup_{a\in\overline{\mathcal{A}}(s)} Q(s,a) < Q_{\max}(s)-\epsilon_{\text{opt}}$, and $Q(s,a) \le Q_{\max}(s)$ on $\mathcal{A}(s)$; hence $\sup_a Q(s,a) = Q_{\max}(s)$. For any $z \in \mathcal{Z}_\mu^*$ and any policy π,

$$Q(s,\pi(s,z)) - \epsilon_{\text{opt}} \le Q_{\max}(s) - \epsilon_{\text{opt}} \le Q(s,\mu_\theta(s,z)), \tag{44}$$

so the indicator vanishes and $\omega_\tau(z) = \tau$ for every $z \in \mathcal{Z}_\mu^*$, independently of π. Since $\omega_\tau(z) > 0$, Eq. (36) of Theorem 2 gives $Q(s,\hat{\pi}^*(s,z)) \ge Q(s,\mu_\theta(s,z)) \ge Q_{\max}(s) - \epsilon_{\text{opt}}$; by conservatism $\hat{\pi}^*(s,z) \notin \overline{\mathcal{A}}(s)$, hence $\hat{\pi}^*(s,z) \in \mathcal{A}_Q^*(s)$ for every $z \in \mathcal{Z}_\mu^*$.

Step 3: Weights on $\mathcal{Z}_\mu^-$. Let $z \in \mathcal{Z}_\mu^- \setminus \mathcal{Z}_{\text{band}}^-(\tau)$, i.e., $Q_{\max}(s) - Q(s,\mu_\theta(s,z)) > \epsilon_{\text{opt}}/c(\alpha_\tau)$. By Eq. (43), $\Delta(\alpha_\tau) > \epsilon_{\text{opt}}$. We check both candidate weights:

- If $\omega_\tau(z) = \tau$, then $\alpha_{\text{reg}}(z) = \alpha_\tau$ and $\Delta(\alpha_\tau) > \epsilon_{\text{opt}}$, so the indicator equals 1 and $\omega_\tau(z) = 1-\tau$, a contradiction.
- If $\omega_\tau(z) = 1-\tau$, then $\alpha_{\text{reg}}(z) = (1-\tau)/\lambda < \alpha_\tau$ since $\tau > 0.5$. By monotonicity, $\Delta((1-\tau)/\lambda) \ge \Delta(\alpha_\tau) > \epsilon_{\text{opt}}$, so the indicator equals 1, which is consistent.

Hence $\omega_\tau(z) = 1-\tau$ is the unique consistent weight on $\mathcal{Z}_\mu^- \setminus \mathcal{Z}_{\text{band}}^-(\tau)$. On the band $\mathcal{Z}_{\text{band}}^-(\tau)$ the improvement may not exceed the margin, and we only use the bound $\omega_\tau(z) \le \tau$.

Step 4: Coupling conditions of Theorem 2. Since $\mathcal{Z}_{\text{test}} = \mathcal{Z}$, $\int_{\mathcal{Z}_{\text{test}}} \delta(z'-z)\,dz' = 1$, and

$$I^* = \int_{\mathcal{Z}_\mu^*} \mathcal{P}_{\mathcal{Z}}(z)\,\omega_\tau(z)\,dz = \tau P^* \ge C, \tag{45}$$

$$\begin{aligned} I^- &= \int_{\mathcal{Z}_\mu^-} \mathcal{P}_{\mathcal{Z}}(z)\,\omega_\tau(z)\,dz \le (1-\tau)\big(P^- - P_{\text{band}}^-(\tau)\big) + \tau P_{\text{band}}^-(\tau) \\ &= (1-\tau)P^- + (2\tau-1)P_{\text{band}}^-(\tau) \le \epsilon. \end{aligned} \tag{46}$$

Since $T(z) = z$, $T(\mathcal{Z}_\mu^*) \cap T(\mathcal{Z}_\mu^-) = \emptyset$ and every fiber is a singleton, so the kernel is non-merging and fiber-wise optimal preservation holds trivially (regime 1 in the proof of Theorem 2). Theorem 2 then yields $\mathbb{P}_{\kappa_Q}(\hat{\pi}^*(s,z) \in \mathcal{A}_Q^*(s)) \ge I^*/(I^*+I^-) \ge \frac{C}{C+\epsilon}$, which proves (a).

Step 5: Inference under the standard prior. At inference, $z \sim \mathcal{N}(0, I_d)$ is not reweighted by ω_τ. By Step 2, every $z \in \mathcal{Z}_\mu^*$ yields an action in $\mathcal{A}_Q^*(s)$, contributing mass P^*. For $z \in \mathcal{Z}_\mu^- \setminus \mathcal{Z}_{\text{band}}^-(\tau)$, Step 3 gives $\omega_\tau(z) = 1-\tau$, and Eq. (41) with $\alpha'_\tau = (1-\tau)/\lambda$ gives $\hat{\pi}^*(s,z) = M(\alpha'_\tau)a^* + (I - M(\alpha'_\tau))\mu_\theta(s,z)$. Define the rectified set

$$\begin{aligned} \mathcal{Z}_{\text{rect}}^-(\tau) &= \Big\{ z \in \mathcal{Z}_\mu^- \setminus \mathcal{Z}_{\text{band}}^-(\tau) : \big\|(I - M(\alpha'_\tau))(\mu_\theta(s,z) - a^*)\big\|_2 \le \sqrt{2\epsilon_{\text{opt}}/L} \Big\}, \\ P_{\text{rect}}^-(\tau) &= \int_{\mathcal{Z}_{\text{rect}}^-(\tau)} \mathcal{P}_{\mathcal{Z}}(z)\,dz. \end{aligned} \tag{47}$$

For $z \in \mathcal{Z}_{\text{rect}}^-(\tau)$, $P \preceq LI$ gives $Q_{\max}(s) - Q(s,\hat{\pi}^*(s,z)) = \frac{1}{2}(\hat{\pi}^* - a^*)^\top P(\hat{\pi}^* - a^*) \le \frac{L}{2}\|\hat{\pi}^* - a^*\|_2^2 \le \epsilon_{\text{opt}}$, and conservatism gives $\hat{\pi}^*(s,z) \in \mathcal{A}_Q^*(s)$. Latents in the band and the remaining latents of $\mathcal{Z}_\mu^-$ are conservatively counted as failures. Therefore

$$\mathbb{P}_{z\sim\mathcal{N}(0,I_d)}\big(\hat{\pi}^*(s,z) \in \mathcal{A}_Q^*(s)\big) \ge P^* + P_{\text{rect}}^-(\tau), \tag{48}$$

which proves (b). Completing the proof. □

A.5 Proof of Proposition 3

Proposition 3. *Suppose the critic Q satisfies the conservatism condition in Theorem 2, and the optimal action set $\mathcal{A}_Q^*(s)$ forms a convex dominant mode over which $Q(s,\cdot)$ is concave. During inference, let $z' \sim \mathcal{N}(0,\sigma^2 I)$ be truncated to $\mathcal{Z}_{test} = \{z' \in \mathcal{Z} \mid \|z'\|_2 \le \sigma\}$. If the truncation radius satisfies $\sigma < \epsilon_{opt}/(Q_{\max}(s) - Q_{\min}(s) + \epsilon_m)$, where ϵ_{opt} is the optimality relaxation factor in Eq. (4), then κ_Q in Eq. (11) fulfills all conditions of Theorem 2 with zero suboptimal leakage ($I^- = 0$), guaranteeing $\mathbb{P}(\hat{\pi}^*(s,z') \in \mathcal{A}_Q^*(s)) = 1$ under the kernel-induced marginal distribution.*

Proof. Recall the definition of the transport mapping for VRD-R in Eq. (11) and Eq. (12). For any $z \in \mathcal{Z} \setminus \{0\}$ (where $\{0\}$ has Lebesgue and Gaussian measure zero), the source latent variable z is mapped to the target coordinate $z_Q(s, z)$:

$$z_Q(s, z) = \Phi_Q(s, z) \cdot \frac{z}{\|z\|_2}, \quad \text{where} \quad \Phi_Q(s, z) = \frac{Q_{\max}(s) - Q(s, \mu_\theta(s, z))}{Q_{\max}(s) - Q_{\min}(s) + \epsilon_m}. \tag{49}$$

Since $Q(s, \mu_\theta(s, z)) \leq Q_{\max}(s)$ holds on $\mathcal{A}(s)$ by definition and on $\bar{\mathcal{A}}(s)$ by the critic conservatism assumption ($\sup_{a \in \bar{\mathcal{A}}(s)} Q(s, a) < Q_{\max}(s) - \epsilon_{\text{opt}}$), we have $\Phi_Q(s, z) \geq 0$ everywhere. Thus, the Euclidean norm of the mapped vector is strictly $\|z_Q(s, z)\|_2 = \Phi_Q(s, z)$, and the alignment kernel is given by $\kappa_Q(z, z' \mid \cdot) = \delta(z' - z_Q(s, z))$.

According to Eq. (4), Definition 1, and the critic conservatism assumption in Theorem 2, for any latent variable $z \in \mathcal{Z}_\mu^- = \mathcal{Z} \setminus \mathcal{Z}_\mu^*$ (which includes both suboptimal in-distribution latents with $\mu_\theta(s, z) \in \mathcal{A}_Q^-(s)$ and OOD latents with $\mu_\theta(s, z) \in \bar{\mathcal{A}}(s)$), the generated behavior action strictly fails the optimality threshold:

$$Q(s, \mu_\theta(s, z)) < Q_{\max}(s) - \epsilon_{\text{opt}}, \quad \forall z \in \mathcal{Z}_\mu^-. \tag{50}$$

Substituting Eq. (50) into the numerator of $\Phi_Q(s, z)$ in Eq. (49), we obtain a strict analytical lower bound for the mapping radius of all non-optimal latents:

$$\|z_Q(s, z)\|_2 = \Phi_Q(s, z) > \frac{Q_{\max}(s) - (Q_{\max}(s) - \epsilon_{\text{opt}})}{Q_{\max}(s) - Q_{\min}(s) + \epsilon_m} = \frac{\epsilon_{\text{opt}}}{Q_{\max}(s) - Q_{\min}(s) + \epsilon_m} \triangleq \delta. \tag{51}$$

This implies that under the geometric reshaping of VRD-R, all suboptimal and OOD behaviors in $\mathcal{Z}_\mu^-$ are strictly repelled to the exterior region with radius $\|z_Q(s, z)\|_2 > \delta$.

Verification of Theorem 2 Condition (i): ϵ-Bounded Suboptimal Leakage. During inference, the test prior $\mathcal{P}_{\text{test}}$ is supported on the closed ball $\mathcal{Z}_{\text{test}} = \{z' \in \mathcal{Z} \mid \|z'\|_2 \leq \sigma\}$. Given the condition $\sigma < \delta$, any $z' \in \mathcal{Z}_{\text{test}}$ satisfies:

$$\|z'\|_2 \leq \sigma < \delta. \tag{52}$$

However, by Eq. (51), every $z \in \mathcal{Z}_\mu^-$ satisfies $\|z_Q(s, z)\|_2 > \delta > \sigma$. Consequently, for all $z \in \mathcal{Z}_\mu^-$, the target image $z_Q(s, z)$ falls strictly outside $\mathcal{Z}_{\text{test}}$, so the Dirac delta measure vanishes identically over $\mathcal{Z}_{\text{test}}$:

$$I^- = \int_{\mathcal{Z}_{\text{test}}} \int_{\mathcal{Z}_\mu^-} \mathcal{P}_{\mathcal{Z}}(z) \kappa_Q(z, z' \mid \cdot)\, dz\, dz' = \int_{\mathcal{Z}_\mu^-} \mathcal{P}_{\mathcal{Z}}(z) \left(\int_{\|z'\|_2 \leq \sigma} \delta(z' - z_Q(s, z))\, dz' \right) dz = 0. \tag{53}$$

Thus, Condition (i) of Theorem 2 ($I^- \leq \epsilon$) holds trivially for any $\epsilon \geq 0$, achieving exact zero suboptimal leakage ($I^- = 0$).

Verification of Theorem 2 Condition (ii): C-Dominant Optimal Coupling. Conversely, for any optimal latent $z \in \mathcal{Z}_\mu^*$, Definition 1 ensures $Q(s, \mu_\theta(s, z)) \geq Q_{\max}(s) - \epsilon_{\text{opt}}$. Substituting this into Eq. (49) yields the upper bound on its mapped radius:

$$\|z_Q(s, z)\|_2 = \Phi_Q(s, z) \leq \delta, \quad \forall z \in \mathcal{Z}_\mu^*. \tag{54}$$

That is, all optimal behaviors are compressed into the closed ball of radius δ centered at the origin, with peak-value actions ($Q(s, \mu_\theta(s, z)) = Q_{\max}(s)$) mapping directly to the origin ($\|z_Q(s, z)\|_2 = 0$). Because the flow policy $\mu_\theta(s, z)$ and critic $Q(s, \cdot)$ are continuous and μ_θ captures the optimal mode where $\max_{z \in \mathcal{Z}} Q(s, \mu_\theta(s, z)) = Q_{\max}(s)$, the pre-image of the test ball $\mathcal{Z}_{\text{test}} = \{z' \mid \|z'\|_2 \leq \sigma\}$ for any $\sigma > 0$ contains a non-empty open neighborhood in $\mathcal{Z}_\mu^*$. Since the Gaussian prior $\mathcal{P}_{\mathcal{Z}}$ has strictly positive density everywhere on $\mathcal{Z}$, integrating over $\mathcal{Z}_{\text{test}}$ yields a strictly positive optimal coupling mass:

$$I^* = \int_{\mathcal{Z}_{\text{test}}} \int_{\mathcal{Z}_\mu^*} \mathcal{P}_{\mathcal{Z}}(z) \delta(z' - z_Q(s, z))\, dz\, dz' = \int_{\{z \in \mathcal{Z}_\mu^* \mid \|z_Q(s,z)\|_2 \leq \sigma\}} \mathcal{P}_{\mathcal{Z}}(z)\, dz \triangleq C > 0. \tag{55}$$

Verification of Mode Separation and Fiber-wise Optimal Preservation. For the transport map $T(z) = z_Q(s, z)$, Eq. (51) shows $T(\mathcal{Z}_\mu^-) \subseteq \{z' \in \mathcal{Z} \mid \|z'\|_2 > \delta\}$, whereas Eq. (54) shows $T(\mathcal{Z}_\mu^*) \subseteq \{z' \in \mathcal{Z} \mid \|z'\|_2 \leq \delta\}$. Therefore, their target images are strictly disjoint:

$$T(\mathcal{Z}_\mu^*) \cap T(\mathcal{Z}_\mu^-) = \emptyset, \tag{56}$$

which verifies the mode separation condition of Theorem 2.

Furthermore, for any $z' \in T(\mathcal{Z}_\mu^*) \cap \mathcal{Z}_{\text{test}}$, its pre-image fiber $T^{-1}(\{z'\}) = \{z \in \mathcal{Z} \mid z_Q(s,z) = z'\}$ is contained entirely within $\mathcal{Z}_\mu^*$. Thus, the induced conditional distribution $\mathcal{W}(\tilde{z} \mid z')$ is supported exclusively on $\mathcal{Z}_\mu^*$, meaning $\mu_\theta(s,\tilde{z}) \in \mathcal{A}_Q^*(s)$ for all $\tilde{z} \in \text{supp}(\mathcal{W}(\cdot \mid z'))$. Since $\mathcal{A}_Q^*(s) \subseteq \mathcal{A}(s)$ is a convex set by assumption, the conditional barycenter $\bar{\mu}(s,z') = \int_{\mathcal{Z}} \mathcal{W}(\tilde{z} \mid z')\mu_\theta(s,\tilde{z})\, d\tilde{z}$ is a convex combination of points in $\mathcal{A}_Q^*(s)$ and therefore remains strictly within the valid in-distribution set: $\bar{\mu}(s,z') \in \mathcal{A}_Q^*(s) \subseteq \mathcal{A}(s)$. Moreover, by the concavity of $Q(s,\cdot)$ on $\mathcal{A}_Q^*(s)$ and Jensen's inequality:

$$Q(s,\bar{\mu}(s,z')) \geq \int_{\mathcal{Z}} \mathcal{W}(\tilde{z} \mid z')Q(s,\mu_\theta(s,\tilde{z}))\, d\tilde{z} \geq Q_{\max}(s) - \epsilon_{\text{opt}}, \tag{57}$$

confirming $\bar{\mu}(s,z') \in \mathcal{A}_Q^*(s)$ and fulfilling the fiber-wise optimal preservation condition required by Eq. (37) in Theorem 2.

Conclusion. Combining $I^- = 0$ (Eq. (53)) and $I^* = C > 0$ (Eq. (55)) with Eq. (33) and Eq. (38) of Theorem 2, we obtain:

$$\mathbb{P}(\hat{\pi}^*(s,z') \in \mathcal{A}_Q^*(s)) \geq \frac{I^*}{I^* + I^-} = \frac{C}{C+0} = 1, \tag{58}$$

which guarantees that under the kernel-induced marginal distribution over $\mathcal{Z}_{\text{test}}$, the VRD-R policy outputs optimal in-distribution actions within $\mathcal{A}_Q^*(s)$ with probability 1, theoretically eliminating OOD Mode Averaging. □

B Algorithms

Algorithm 1 Quantile-based Value-Rectified Distillation (VRD-Q) Algorithm

Require: Pre-trained flow policy μ_θ, Q-function evaluator Q, static dataset $\mathcal{D}$, base prior $\mathcal{Z}$, target actor policy π

1: **for** each batch $s \sim \mathcal{D}$, $z \sim \mathcal{Z}$ **do**
2: **1.** Calculate quantile-aware weight:
$\omega_\tau(z) = |\tau - \mathbb{I}(Q(s,\mu_\theta(s,z)) < Q(s,\pi(s,z)) - \epsilon_{\text{opt}})|$
3: **2.** Construct in-place alignment kernel:
$\kappa_Q(z,z'|s,\mu_\theta,\pi) = \omega_\tau(z) \cdot \delta(z'-z)$
4: **3.** Minimize Quantile-Rectified Loss to update π:
$\mathcal{L}_\pi = \mathbb{E}\left[\omega_\tau(z) \cdot ||\mu_\theta(s,z) - \pi(s,z)||_2^2 - \lambda \cdot \mathbb{E}_{z' \sim \mathcal{P}_{test}}[Q(s,\pi(s,z'))]\right]$
5: **end for**
6: **Zero-Shot Inference (VRD-Q):** Sample $z \sim \mathcal{N}(0, I_d)$, output $a = \pi^*(s,z)$

Algorithm 2 Reshaping-based Value-Rectified Distillation (VRD-R) Algorithm

Require: Pre-trained flow policy μ_θ, Q-function evaluator Q, static dataset $\mathcal{D}$, base prior $\mathcal{Z}$, target actor policy π

1: **for** each batch $s \sim \mathcal{D}$, $z \sim \mathcal{Z}$ **do**
2: **1.** Calculate normalized ranking scalar:
$\Phi_Q(s,z) = \frac{Q_{max}(s) - Q(s,\mu_\theta(s,z))}{Q_{max}(s) - Q_{min}(s) + \epsilon_m}$
3: **2.** Apply value-guided reshaping function (gravitational field towards origin):
$z_Q(s,z,\mu_\theta) = \Phi_Q(s,z) \cdot \frac{z}{||z||_2}$
4: **3.** Construct transport alignment kernel:
$\kappa_Q(z,z'|s,\mu_\theta,\pi) = 1 \cdot \delta(z' - z_Q(s,z,\mu_\theta))$
5: **4.** Minimize Reshaping-Rectified Loss to update π:
$\mathcal{L}_\pi = \mathbb{E}\left[||\mu_\theta(s,z) - \pi(s,z_Q(s,z,\mu_\theta))||_2^2 - \lambda \cdot \mathbb{E}_{z' \sim \mathcal{P}_{test}}[Q(s,\pi(s,z'))]\right]$
6: **end for**
7: **Zero-Shot Inference (VRD-R):** Sample $z' \sim \tilde{\mathcal{N}}(0, \sigma^2 I_d)$ (where $\sigma \ll 1$), output $a = \pi^*(s,z')$

C External experimental details

C.1 Full results over OGBenchmarks (50 tasks)

Dataset Overview. OGBench (Offline Goal-Conditioned RL Benchmark) is a large-scale evaluation suite comprising 50 tasks across 10 diverse domains, including complex navigation (e.g., *antmaze-giant*, *humanoidmaze*), robotic manipulation (*cube-play*, *scene-play*), and logical puzzles (*puzzle-3x3/4x4*). These tasks are specifically designed to test an agent's ability to reach diverse goals using multi-modal offline data, presenting significant challenges in mode separation and long-horizon planning.

Table 4: Detailed Offline RL Results on OGBench Tasks: Performance comparison across different offline RLalgorithms. All prior work results are taken from [21] to compare a wide range of previous methods against VRD.

Task	BC	IQL	ReBRAC	IDQL	SRPO	CAC	FAWAC	FBRAC	IFQL	FQL	GFP actor π_ω	**VRD-Q (Ours)**	**VRD-R (Ours)**
antmaze-large-navigate-singletask													
task1-v0	0 ± 0	48 ± 9	91 ± 10	0 ± 0	0 ± 0	42 ± 7	1 ± 1	70 ± 20	24 ± 17	80 ± 8	95.4 ± 0.8	$\mathbf{96.7} \pm 0.4$	$\mathbf{97.2} \pm 0.6$
task2-v0	6 ± 3	42 ± 6	88 ± 4	14 ± 8	4 ± 4	1 ± 1	0 ± 1	35 ± 12	8 ± 3	57 ± 10	92.2 ± 3.0	$\mathbf{96.2} \pm 0.1$	93.1 ± 3.3
task3-v0	29 ± 5	72 ± 7	51 ± 18	26 ± 8	3 ± 2	49 ± 10	12 ± 4	83 ± 15	52 ± 17	93 ± 3	95.6 ± 2.7	$\mathbf{98.4} \pm 0.7$	96.4 ± 1.7
task4-v0	8 ± 3	51 ± 9	84 ± 7	62 ± 25	45 ± 19	17 ± 6	10 ± 3	37 ± 18	18 ± 8	80 ± 4	$\mathbf{90.6} \pm 2.6$	$\mathbf{90.7} \pm 3.8$	$\mathbf{90.3} \pm 2.9$
task5-v0	10 ± 3	54 ± 22	90 ± 2	2 ± 2	1 ± 1	55 ± 6	9 ± 5	76 ± 8	38 ± 18	83 ± 4	95.0 ± 1.3	95.4 ± 0.9	$\mathbf{97.5} \pm 0.8$
antmaze-giant-navigate-singletask-task													
task1-v0	0 ± 0	0 ± 0	27 ± 22	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 1	0 ± 0	4 ± 5	12.6 ± 15.1	$\mathbf{62.4} \pm 1.7$	6.4 ± 2.6
task2-v0	0 ± 0	1 ± 1	16 ± 17	0 ± 0	0 ± 0	0 ± 0	0 ± 0	4 ± 7	0 ± 0	9 ± 7	52.2 ± 26.5	$\mathbf{94.2} \pm 0.4$	87.0 ± 2.6
task3-v0	0 ± 0	0 ± 0	$\mathbf{34} \pm 22$	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 1	13.7 ± 10.2	28.6 ± 0.4	3.5 ± 2.2
task4-v0	0 ± 0	0 ± 0	5 ± 12	0 ± 0	0 ± 0	0 ± 0	0 ± 0	9 ± 4	0 ± 0	14 ± 23	$\mathbf{17.8} \pm 19.9$	6.4 ± 8.5	14.3 ± 15.8
task5-v0	1 ± 1	19 ± 7	49 ± 22	0 ± 1	0 ± 0	0 ± 0	0 ± 0	6 ± 10	13 ± 9	16 ± 28	43.2 ± 35.7	57.3 ± 15.5	$\mathbf{80.3} \pm 12.7$
humanoidmaze-medium-navigate-singletask													
task1-v0	1 ± 0	32 ± 7	16 ± 9	1 ± 1	0 ± 0	38 ± 19	6 ± 2	25 ± 8	69 ± 19	19 ± 12	83.5 ± 3.7	$\mathbf{87.2} \pm 2.8$	$\mathbf{85.1} \pm 6.6$
task2-v0	1 ± 0	41 ± 9	18 ± 16	1 ± 1	1 ± 1	47 ± 35	40 ± 2	76 ± 10	85 ± 11	$\mathbf{94} \pm 3$	91.2 ± 6.3	$\mathbf{95.8} \pm 1.6$	91.7 ± 0.8
task3-v0	6 ± 2	25 ± 5	36 ± 13	0 ± 1	2 ± 1	83 ± 18	19 ± 2	27 ± 11	49 ± 49	74 ± 18	86.3 ± 10.7	$\mathbf{92.7} \pm 4.3$	78.4 ± 19.6
task4-v0	0 ± 0	0 ± 1	15 ± 16	1 ± 1	1 ± 1	5 ± 4	1 ± 1	1 ± 2	1 ± 1	3 ± 4	3.0 ± 6.6	4.7 ± 4.9	2.1 ± 2.5
task5-v0	2 ± 1	66 ± 4	24 ± 20	1 ± 1	3 ± 3	91 ± 5	31 ± 7	63 ± 9	$\mathbf{98} \pm 2$	$\mathbf{97} \pm 2$	$\mathbf{95.8} \pm 1.3$	$\mathbf{96.3} \pm 0.7$	94.1 ± 3.8
humanoidmaze-large-navigate-singletask													
task1-v0	0 ± 0	3 ± 1	2 ± 1	0 ± 0	0 ± 0	1 ± 1	0 ± 0	0 ± 1	6 ± 2	7 ± 6	57.2 ± 23.7	$\mathbf{68.2} \pm 12.4$	39.1 ± 18.4
task2-v0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	$\mathbf{0.1} \pm 0.2$	$\mathbf{0.1} \pm 0.2$	$\mathbf{0.8} \pm 1.1$
task3-v0	1 ± 1	7 ± 3	8 ± 4	3 ± 1	1 ± 1	2 ± 3	1 ± 1	10 ± 2	$\mathbf{48} \pm 10$	11 ± 7	14.6 ± 16.2	12.7 ± 15.1	37.8 ± 6.1
task4-v0	1 ± 0	1 ± 0	1 ± 1	0 ± 0	0 ± 0	0 ± 1	0 ± 0	0 ± 0	1 ± 1	2 ± 3	3.7 ± 4.1	$\mathbf{4.7} \pm 4.4$	1.7 ± 1.4
task5-v0	0 ± 1	1 ± 1	2 ± 2	0 ± 0	0 ± 0	0 ± 0	0 ± 0	1 ± 1	0 ± 0	1 ± 3	$\mathbf{13.1} \pm 15.5$	$\mathbf{12.9} \pm 14.2$	4.4 ± 2.6
antsoccer-arena-navigate-singletask													
task1-v0	2 ± 1	14 ± 5	0 ± 0	44 ± 12	2 ± 1	1 ± 3	22 ± 2	17 ± 3	61 ± 25	77 ± 4	77.0 ± 1.7	$\mathbf{82.3} \pm 2.1$	52.5 ± 11.2
task2-v0	2 ± 2	17 ± 7	0 ± 1	15 ± 12	3 ± 1	0 ± 0	8 ± 1	8 ± 2	75 ± 3	$\mathbf{88} \pm 3$	$\mathbf{91.2} \pm 2.2$	$\mathbf{91.3} \pm 2.4$	78.7 ± 2.8
task3-v0	0 ± 0	6 ± 4	0 ± 0	0 ± 0	0 ± 0	8 ± 19	11 ± 5	16 ± 3	14 ± 22	$\mathbf{61} \pm 6$	51.9 ± 4.9	$\mathbf{59.8} \pm 5.8$	52.4 ± 12.7
task4-v0	1 ± 0	3 ± 2	0 ± 0	0 ± 1	0 ± 0	0 ± 0	12 ± 3	24 ± 4	16 ± 9	39 ± 6	40.2 ± 4.2	$\mathbf{44.1} \pm 5.6$	22.7 ± 14.0
task5-v0	0 ± 0	2 ± 2	0 ± 0	0 ± 0	0 ± 0	0 ± 0	9 ± 2	15 ± 4	0 ± 1	36 ± 9	29.1 ± 8.9	$\mathbf{41.7} \pm 7.4$	33.3 ± 7.6
cube-single-play-singletask													
task1-v0	10 ± 5	88 ± 3	89 ± 5	$\mathbf{95} \pm 2$	89 ± 7	77 ± 28	81 ± 9	73 ± 33	79 ± 4	$\mathbf{97} \pm 2$	$\mathbf{99.1} \pm 0.4$	$\mathbf{100} \pm 0$	$\mathbf{100} \pm 0$
task2-v0	3 ± 1	85 ± 8	92 ± 4	$\mathbf{96} \pm 2$	82 ± 16	80 ± 30	81 ± 9	83 ± 13	73 ± 3	$\mathbf{97} \pm 2$	$\mathbf{99.4} \pm 0.7$	$\mathbf{100} \pm 0$	97.5 ± 1.2
task3-v0	9 ± 3	91 ± 5	93 ± 3	$\mathbf{99} \pm 1$	$\mathbf{96} \pm 2$	$\mathbf{98} \pm 1$	87 ± 4	82 ± 12	88 ± 4	$\mathbf{98} \pm 2$	$\mathbf{99.4} \pm 0.5$	$\mathbf{100} \pm 0$	$\mathbf{100} \pm 0$
task4-v0	2 ± 1	73 ± 6	92 ± 3	93 ± 4	70 ± 18	91 ± 2	79 ± 6	79 ± 20	79 ± 6	94 ± 3	$\mathbf{99.1} \pm 0.7$	$\mathbf{99.1} \pm 0.4$	92.7 ± 2.4
task5-v0	3 ± 3	78 ± 9	87 ± 8	90 ± 6	61 ± 12	80 ± 20	78 ± 10	76 ± 33	77 ± 7	$\mathbf{93} \pm 3$	$\mathbf{97.0} \pm 1.6$	$\mathbf{98.2} \pm 2.4$	$\mathbf{98.4} \pm 0.2$
cube-double-play-singletask													
task1-v0	8 ± 3	27 ± 5	45 ± 6	39 ± 19	7 ± 6	21 ± 8	21 ± 7	47 ± 11	35 ± 9	61 ± 9	$\mathbf{76.1} \pm 4.6$	72.4 ± 3.1	63.6 ± 8.1
task2-v0	0 ± 0	1 ± 1	7 ± 3	16 ± 10	0 ± 0	2 ± 2	2 ± 1	22 ± 12	9 ± 5	36 ± 6	$\mathbf{53.3} \pm 8.9$	48.4 ± 10.3	17.4 ± 7.1
task3-v0	0 ± 0	0 ± 0	4 ± 1	17 ± 8	0 ± 1	3 ± 1	1 ± 1	4 ± 2	8 ± 5	22 ± 5	$\mathbf{43.3} \pm 8.9$	$\mathbf{44.4} \pm 5.1$	24.7 ± 14.4
task4-v0	0 ± 0	0 ± 0	1 ± 1	0 ± 1	0 ± 0	0 ± 1	0 ± 0	0 ± 1	1 ± 1	5 ± 2	$\mathbf{7.1} \pm 3.1$	$\mathbf{7.0} \pm 2.7$	6.5 ± 3.3
task5-v0	0 ± 0	4 ± 3	4 ± 2	1 ± 1	0 ± 0	3 ± 2	2 ± 1	2 ± 2	17 ± 6	19 ± 10	$\mathbf{56.3} \pm 11.3$	$\mathbf{55.7} \pm 10.7$	48.9 ± 9.3
scene-play-singletask													
task1-v0	19 ± 6	94 ± 3	$\mathbf{95} \pm 2$	$\mathbf{100} \pm 0$	94 ± 4	$\mathbf{100} \pm 1$	87 ± 8	$\mathbf{96} \pm 8$	$\mathbf{98} \pm 3$	$\mathbf{100} \pm 0$	$\mathbf{99.8} \pm 0.4$	$\mathbf{100} \pm 0$	$\mathbf{100} \pm 0$
task2-v0	1 ± 1	12 ± 3	50 ± 13	33 ± 14	2 ± 2	50 ± 40	18 ± 8	46 ± 10	0 ± 0	76 ± 9	89.0 ± 4.1	$\mathbf{99.2} \pm 0.6$	$\mathbf{97.3} \pm 1.7$
task3-v0	1 ± 1	32 ± 7	55 ± 16	$\mathbf{94} \pm 4$	4 ± 4	49 ± 16	38 ± 9	78 ± 14	54 ± 19	$\mathbf{98} \pm 1$	78.0 ± 13.2	75.4 ± 14.5	81.4 ± 10.4
task4-v0	2 ± 2	0 ± 1	3 ± 3	4 ± 3	0 ± 0	0 ± 0	$\mathbf{6} \pm 1$	4 ± 4	0 ± 0	5 ± 1	0.6 ± 0.6	2.4 ± 1.1	0.4 ± 0.3
task5-v0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
puzzle-3x3-play-singletask													
task1-v0	5 ± 2	33 ± 6	97 ± 4	52 ± 12	89 ± 5	97 ± 2	25 ± 9	63 ± 19	94 ± 3	90 ± 4	94.8 ± 4.4	$\mathbf{99.0} \pm 0.4$	$\mathbf{97.6} \pm 2.1$
task2-v0	1 ± 1	4 ± 3	1 ± 1	0 ± 1	0 ± 1	0 ± 0	4 ± 2	2 ± 2	1 ± 2	16 ± 5	0.3 ± 0.3	7.7 ± 5.3	$\mathbf{20.3} \pm 22.6$
task3-v0	1 ± 1	3 ± 2	3 ± 1	0 ± 0	0 ± 0	0 ± 0	1 ± 0	1 ± 1	0 ± 0	$\mathbf{10} \pm 3$	0.9 ± 0.6	4.7 ± 4.2	1.0 ± 0.4
task4-v0	1 ± 1	2 ± 1	2 ± 1	0 ± 0	0 ± 0	0 ± 0	1 ± 1	2 ± 2	0 ± 0	16 ± 5	5.4 ± 2.1	16.2 ± 4.3	$\mathbf{34.1} \pm 30.4$
task5-v0	1 ± 0	3 ± 2	5 ± 3	0 ± 0	0 ± 0	0 ± 0	1 ± 1	2 ± 2	0 ± 0	16 ± 3	14.1 ± 8.4	15.7 ± 4.6	$\mathbf{20.7} \pm 2.6$
puzzle-4x4-play-singletask													
task1-v0	1 ± 1	12 ± 2	26 ± 4	48 ± 5	24 ± 9	44 ± 10	1 ± 2	32 ± 9	49 ± 9	34 ± 8	50.0 ± 8.1	$\mathbf{58.7} \pm 5.6$	35.4 ± 12.7
task2-v0	0 ± 0	7 ± 4	12 ± 4	14 ± 5	0 ± 1	0 ± 0	0 ± 1	5 ± 3	4 ± 4	$\mathbf{16} \pm 5$	9.9 ± 2.5	$\mathbf{15.6} \pm 1.7$	6.7 ± 3.4
task3-v0	0 ± 0	9 ± 3	15 ± 3	34 ± 5	21 ± 10	29 ± 12	1 ± 1	20 ± 10	$\mathbf{50} \pm 14$	18 ± 5	46.2 ± 3.8	$\mathbf{49.8} \pm 2.9$	24.4 ± 9.8
task4-v0	0 ± 0	5 ± 2	10 ± 3	$\mathbf{26} \pm 6$	7 ± 4	1 ± 1	0 ± 0	5 ± 1	21 ± 11	11 ± 3	17.2 ± 2.5	20.4 ± 3.4	$\mathbf{25.7} \pm 2.1$
task5-v0	0 ± 0	4 ± 1	7 ± 3	$\mathbf{24} \pm 11$	1 ± 1	0 ± 0	0 ± 1	4 ± 3	2 ± 2	7 ± 3	7.3 ± 3.6	8.6 ± 4.1	6.7 ± 3.4
Average (50 tasks)	2.8	23.4	31.0	22.8	14.3	25.3	16.1	28.6	30.0	43.5	51.8	**56.4**	50.8

Results and Analysis. As shown in Table 4, our proposed VRD framework demonstrates superior performance across the vast majority of the 50 tasks. Specifically:

- **Overall Superiority**: VRD-Q achieves the highest overall average score of **56.4**, significantly outperforming the strongest generative baseline, GFP actor (51.8), and the state-of-the-art FQL (43.5). Besides, VRD-R achieves competitive average score of 50.8.
- **Robustness in Challenging Domains**: In highly complex environments like *antmaze-giant* and *humanoidmaze*, where conventional baselines such as IQL, ReBRAC, and IDQL often struggle or yield near-zero scores, VRD-Q and VRD-R maintain high success rates (e.g., 94.2% in *antmaze-giant* task2). This highlights the effectiveness of our value-rectification mechanism in navigating sparse-reward, high-dimensional spaces.
- **High-Precision Manipulation**: In manipulation tasks like *cube-single-play* and *scene-play*, VRD-Q reaches a 100% success rate over some tasks, demonstrating its capability to precisely align generative trajectories with optimal expert behaviors.
- **Variant Comparison**: While VRD-Q is generally more stable and leads in average performance, VRD-R shows competitive advantages in specific structured tasks, such as *puzzle-3x3* and *antsoccer*, suggesting that geometric latent reshaping is particularly effective for certain task dynamics.

In conclusion, these results validate that VRD effectively addresses the latent fusion problem, providing a more robust and high-performing solution for goal-conditioned offline reinforcement learning.

C.2 Minari: External comparison study

The MuJoCo task suite within the Minari dataset [26] serves as a standardized benchmark widely adopted in offline reinforcement learning (RL) research. Developed by the Farama Foundation as a modern successor to the D4RL dataset, these tasks are built upon classic physics simulation environments, including *HalfCheetah*, *Hopper*, and *Walker2d*. Minari provides a tiered classification system to evaluate algorithms across varying data qualities:

- **Simple Datasets**: These are primarily generated by random or early-stage training policies. They contain a high volume of low-value, exploratory trajectories characterized by a divergent distribution and a low signal-to-noise ratio, testing an algorithm's ability to extract fundamental control logic from poor-quality data.
- **Medium Datasets**: Collected from policies at intermediate performance levels, these datasets are goal-oriented but frequently contain sub-optimal actions and failed attempts. They serve as a core benchmark for evaluating whether an algorithm can effectively isolate optimal modes and outperform the data average.
- **Expert Datasets**: Generated by fully converged, high-performance policies, these represent optimal behavior with a narrow, high-quality distribution. They are designed to verify if an algorithm can achieve high-precision policy alignment and maintain expert-level performance under limited data diversity.

By adhering to the Gymnasium API and providing comprehensive metadata, Minari serves as an ideal testbed for evaluating capacity for mode separation and distributional correction within complex, multi-modal, and non-expert datasets.

Flow Q-Learning (FQL) [21] and Guided Flow Policy (GFP) [24] are selected as the main baselines for comparison study.

Results Overview. Table 5 summarizes the performance of our proposed VRD framework against several state-of-the-art generative baselines across various task levels in the Minari benchmark. Overall, VRD-R achieves the highest average score of **112.8**, outperforming both the flow-based FQL (99.7) and the strongest GFP variants (108.4).

Key Observations.

- **Consistent Superiority in Complex Environments**: In the *Walker2d* and *Hopper* domains, VRD-Q and VRD-R consistently outperform the baselines. Notably, in *hopper-simple*, both VRD variants achieve a score of approximately **99.1**, nearly doubling the performance of FQL (57.2), which demonstrates the robustness of our framework in extracting optimal policies from low-quality data.

Table 5: Comparisons on Minari MuJoCo benchmarks. The best scores are highlighted in bold.

Task	Offline RL algorithms				
	FQL	**GFP** actor π_θ	**GFP** VaBC π_ω	**VRD-Q (Ours)**	**VRD-R (Ours)**
Minari halfcheetah-simple-v0	59.2 ± 0.3	**72.5** ± 0.5	64.4 ± 0.4	71.4 ± 0.1	70.3 ± 0.2
Minari halfcheetah-medium-v0	100.2 ± 6.7	121.3 ± 5.8	108.6 ± 5.3	117.4 ± 3.1	**123.9** ± 2.3
Minari halfcheetah-expert-v0	134.1 ± 2.3	133.4 ± 1.3	136.4 ± 0.8	96.7 ± 10.7	**139.0** ± 0.1
Minari hopper-simple-v0	57.2 ± 5.9	91.6 ± 4.3	87.4 ± 6.6	**99.1** ± 0.1	**99.2** ± 0.6
Minari hopper-medium-v0	81.9 ± 23.9	79.6 ± 14.5	78.2 ± 24.2	**107.3** ± 4.4	104.3 ± 8.6
Minari hopper-expert-v0	99.6 ± 10.9	103.9 ± 10.6	**108.8** ± 13.1	107.3 ± 9.1	101.4 ± 10.2
Minari walker2d-simple-v0	89.4 ± 0.7	90.0 ± 0.9	89.7 ± 0.9	**91.5** ± 0.2	**91.5** ± 0.7
Minari walker2d-medium-v0	127.6 ± 3.0	133.7 ± 1.1	126.1 ± 3.5	**135.6** ± 0.3	134.0 ± 1.7
Minari walker2d-expert-v0	148.0 ± 1.8	149.9 ± 2.0	150.8 ± 0.4	**152.3** ± 0.2	151.4 ± 0.1
Average	99.7	108.4	105.6	108.7	**112.8**

- **High-Value Mode Recovery**: VRD-R exhibits exceptional performance in *halfcheetah-expert*, reaching a peak score of **139.0**. This suggests that the geometric latent reshaping in VRD-R is highly effective at aligning generative trajectories with expert behaviors in high-dimensional action spaces.
- **Mitigation of Mode Averaging**: The significant performance gains observed in *Simple* and *Medium* datasets—which typically contain a higher proportion of sub-optimal trajectories—validate that our value-rectification mechanism successfully mitigates the "latent fusion" problem. By decoupling optimal modes from sub-optimal noise, VRD maintains high performance where standard generative models often suffer from catastrophic mode averaging.

In conclusion, the results across Minari tasks confirm that the VRD framework is not only competitive in expert-level tasks but also exceptionally robust in handling non-expert, multi-modal data.

C.3 Full sensitivity analysis for hyperparameters

To provide a comprehensive understanding of the stability and robustness of the Value-Rectified Distillation (VRD) framework, we conduct an extensive grid search on two primary hyperparameters: the quantile threshold τ for VRD-Q and the truncation radius σ for VRD-R, alongside the balance coefficient λ. The results are visualized in Figure 5 and Figure 6.

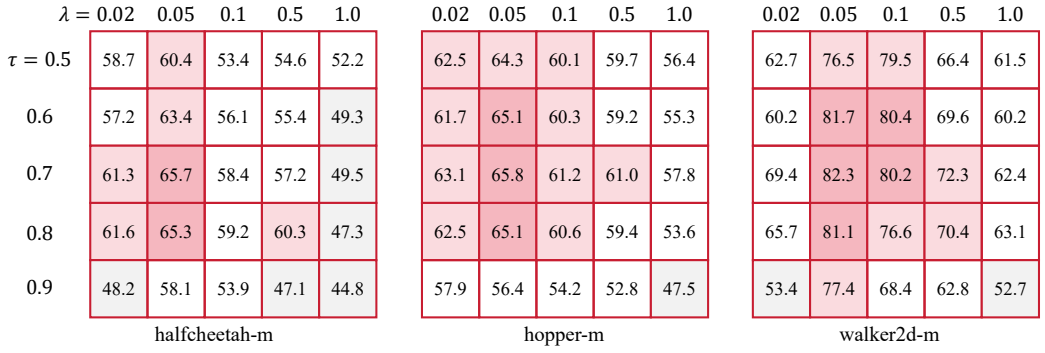


Figure 5: Sensitivity analysis for VRD-Q.

C.3.1 Analysis of VRD-Q Sensitivity (Figure 5)

The experimental results on D4RL ‘medium‘ datasets (HalfCheetah, Hopper, and Walker2d) reveal several key insights for VRD-Q:

- **Inverted U-shape performance w.r.t. τ:** The performance generally follows an inverted U-shaped curve across all environments. When τ is too small (e.g., 0.5), the statistical filtering is insufficient to suppress sub-optimal behavioral modes. Conversely, an excessively large τ (e.g., 0.9) overly restricts the distillation weight, leading to a loss of distribution coverage and potential vanishing gradients.
- **Optimal τ Region:** The peak performance consistently resides in the range of $\tau \in [0.7, 0.8]$. This range effectively balances the exclusion of low-value "OOD noise" with the preservation of high-value multi-modal trajectories.
- **Robustness to λ:** VRD-Q demonstrates relative stability across a wide range of λ (from 0.02 to 0.1). However, very large values of λ (e.g., 1.0) tend to degrade performance as the rectification term begins to dominate the Q-value maximization objective.

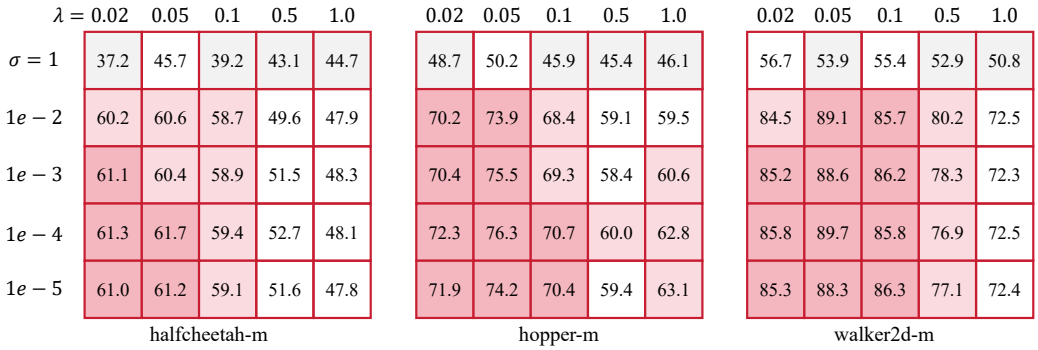


Figure 6: Sensitivity analysis for VRD-R.

C.3.2 Analysis of VRD-R Sensitivity (Figure 6)

For the geometric rectification variant VRD-R, the sensitivity analysis highlights:

- **Criticality of Truncation Radius σ:** Performance is highly sensitive to the scale of the latent prior during inference. When $\sigma = 1$ (equivalent to standard Gaussian sampling), the performance is sub-optimal because the sampler hits the peripheral sub-optimal regions.
- **Optimal σ Scale:** As σ decreases towards 10^{-4}, the performance significantly improves and stabilizes. This confirms that the value-guided gravity field successfully compresses high-value actions toward the latent origin, and a smaller σ ensures the sampler stays within this "expert core."
- **Environment Consistency:** Similar to VRD-Q, the best configuration for VRD-R is relatively consistent across different locomotion tasks, with $\lambda = 0.05$ and $\sigma = 10^{-4}$ providing the most robust results.

C.3.3 Practical Hyperparameter Tuning Suggestions

Based on our empirical analysis, we offer the following guidelines for applying VRD to new offline RL tasks:

1. **Tuning the Quantile Threshold τ (VRD-Q):**
   - **Default:** Start with $\tau = 0.7$.
   - **High-Noise / Exploratory Datasets:** While Theorem 2 suggests a larger τ under an oracle critic, on noisy datasets (e.g., “random” or “medium-replay”) the learned Q-network inherently suffers from higher estimation variance. Setting τ too close to 1.0 makes the suboptimal anchor weight $1 - \tau$ vanishingly small, causing the actor to exploit spurious OOD Q-peaks and destabilizing training. Thus, we recommend decreasing τ slightly (e.g., to $0.6 \sim 0.7$) to retain a necessary behavioral regularization

margin ($1 - \tau \in [0.3, 0.4]$) for training stability while still down-weighting suboptimal modes.

- **Expert / Structured Datasets:** For datasets with high-quality demonstrations where value estimation is accurate and stable, a larger τ (e.g., $0.8 \sim 0.9$) can be safely employed to aggressively concentrate the policy onto the peak optimal trajectories without risking OOD divergence.

2. **Tuning the Truncation Radius σ (VRD-R):**
   - **Default:** Set $\sigma = 10^{-4}$.
   - **Geometric Adjustment:** If the agent exhibits "mode collapse" (too little variety in actions), increase σ to 10^{-3}. If the agent still chooses sub-optimal actions in multi-modal settings, decrease σ to 10^{-5} to force more aggressive centering.
3. **Tuning the Balance Coefficient λ:**
   - We recommend $\lambda \in [0.05, 0.1]$. A smaller λ (e.g., 0.02) is preferred when the Q-function is well-trained and accurate. If the environment is complex and requires stronger regularization to prevent OOD mode-averaging, increase λ to 0.5.

C.4 Validation study

C.4.1 Quantile Distillation in VRD-Q

To qualitatively investigate the efficacy of the proposed Value-Rectified Distillation (VRD-Q) in isolating high-value behaviors, we conduct a visualization analysis of the latent space. Specifically, we apply t-SNE to reduce the dimensionality of the standard prior noise vectors z, which are sampled during policy inference, and project them into a 2D plane. We evaluate three representative tasks from the D4RL benchmark: `halfcheetah-medium-v2`, `hopper-medium-v2`, and `walker2d-medium-v2`. In the visualization, each point represents a sampled latent code z, color-coded to indicate whether the action generated by the VRD-Q Actor or the baseline Flow Model policy yields a higher estimated Q-value.

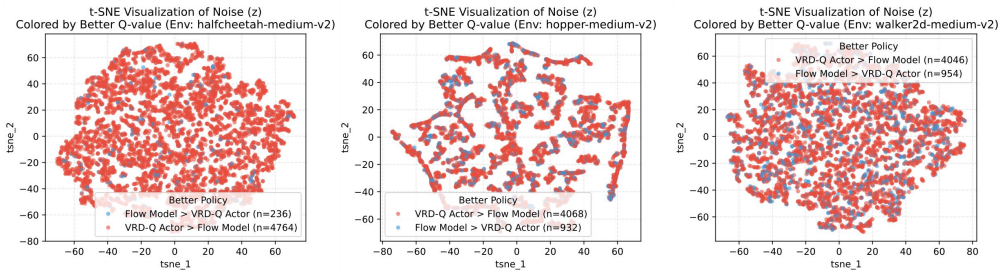


Figure 7: Validation study for Quantile Distillation in VRD-Q.

The experimental results, as illustrated in Figure 7, reveal that the VRD-Q Actor consistently generates actions with superior Q-values across the vast majority of the latent space compared to the original Flow Model. The quantitative distribution of "better" samples is summarized as follows:

- **Halfcheetah-medium-v2**: The VRD-Q Actor produces higher-value actions for $4,764$ samples, while the Flow Model only excels in 236 samples.
- **Hopper-medium-v2**: The VRD-Q Actor outperforms the baseline in $4,068$ samples, significantly exceeding the 932 samples of the Flow Model.
- **Walker2d-medium-v2**: Higher Q-values are achieved by the VRD-Q Actor in $4,046$ samples, compared to 954 for the Flow Model.

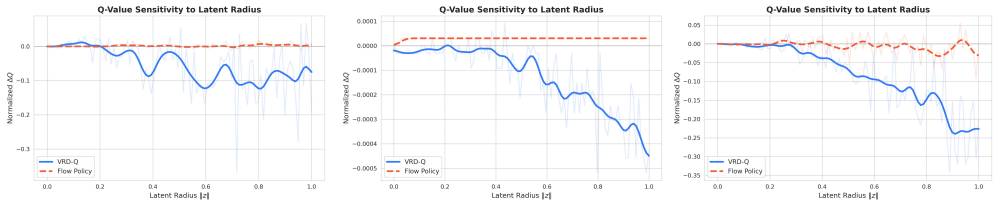
This visualization provides strong empirical evidence for the rationality of the VRD-Q framework. As theoretically analyzed, standard Flow Matching policies often suffer from *Latent Fusion* in multi-modal offline datasets, where optimal and sub-optimal generative trajectories become topologically

entangled. Direct Q-guided distillation without rectification risks *Out-of-Distribution (OOD) Mode Averaging*, leading to policy collapse into low-value regions. In contrast, VRD-Q utilizes a *quantile-aware weight modulation* mechanism to statistically filter out sub-optimal behaviors while maintaining the structural integrity of the latent space. The fact that the majority of the prior noise distribution is mapped to higher-value actions demonstrates that VRD-Q effectively decouples behavioral modes and successfully rectifies the flow dynamics toward the high-value action manifold, thereby ensuring robust and high-performance policy execution.

C.5 Reshaping Distillation in VRD-R

To investigate the effectiveness of the proposed VRD-R in achieving geometric separation of behavioral modes, we conduct a sensitivity analysis of Q-values relative to the latent space scale.

Experimental Setup. We evaluate three representative tasks from the D4RL benchmark: `halfcheetah-medium-v2`, `hopper-medium-v2`, and `walker2d-medium-v2`. The horizontal axis represents the *Latent Radius* $\|z\|$, which denotes the magnitude of the standard prior noise sampled during the inference process, ranging from 0.0 to 1.0. The vertical axis measures the *Normalized* ΔQ, representing the normalized change in the estimated Q-values of actions generated at the corresponding latent radius. This setup allows us to track the topological distribution of action values across the latent manifold.



halfcheetah-m hopper-m walker2d-m

Figure 8: Validation for VRD-R: Q-value sensitivity to latent radius

Results and Rationality Analysis. As illustrated in Fig. 8, the experimental results provide strong empirical evidence for the geometric reorganization capabilities of VRD-R:

- **Intentional Value Decay Gradient:** Across all tested environments, the VRD-R curves exhibit a significant and consistent downward trend as the latent radius $\|z\|$ increases. This sharp decline in Normalized ΔQ is a key indicator of successful latent reshaping.
- **Spatial Isolation of Modes:** The observed gradient confirms that VRD-R has successfully decoupled the intertwined modes found in the original multi-modal datasets. Specifically, it forces the *optimal modes* (high-value behaviors) to consolidate within the central region of the latent space (near the origin, $\|z\| \approx 0$), while simultaneously pushing *sub-optimal modes* (low-value behaviors) toward the peripheral, high-variance regions (larger $\|z\|$).
- **Strategic Alignment with Gaussian Prior:** Given that the standard Gaussian prior utilized during inference naturally assigns the highest probability density to the region around the origin, this "center-optimal, edge-suboptimal" topology is highly rational. It ensures that standard sampling is highly likely to hit the high-value action manifold.

In conclusion, the sensitivity results demonstrate that VRD-R effectively rectifies the flow dynamics by geometrically isolating high-value trajectories. This mechanism fundamentally prevents the *Mode Averaging* issue by ensuring that sub-optimal behaviors are spatially excluded from the primary inference region, thereby significantly enhancing the policy's performance and robustness.

C.6 Hyperparameters

In this section, we detail the hyperparameters used for evaluating our proposed Value-Rectified Distillation (VRD) framework, including both the VRD-Q and VRD-R instantiations. To ensure

a fair and rigorous comparison, the foundational hyperparameters (e.g., network architectures, learning rates, and flow matching configurations) are kept consistently identical across all evaluated benchmarks, namely D4RL, OGBench, and Minari. Task-specific tuning is strictly limited to the core hyperparameters of the VRD framework (τ for VRD-Q and σ for VRD-R) to adapt to the varying degrees of multi-modality, dataset quality, and environment complexity.

General Network and Training Configurations Both the value networks (critics) and the flow-based policy networks (actors) are parameterized using Multi-Layer Perceptrons (MLPs). For the flow matching process, we maintain a fixed number of integration steps (NFE = 10) using the Euler ODE solver. The general architectural choices and optimization hyperparameters are summarized in Table 6.

VRD-Specific Hyperparameters and Task Dependency The two algorithmic variants of our framework require specific hyperparameter configurations to effectively decouple sub-optimal and optimal behavioral modes. Empirically, we observe a strong dependency between the dataset profile (i.e., multi-modality and sub-optimality) and the optimal hyperparameter selection:

- **VRD-Q (Quantile-based Filtering):** The primary hyperparameter is the quantile threshold $\tau \in (0.5, 1)$, which assigns a smaller distillation weight $1 - \tau$ to suboptimal trajectories and a larger weight τ to optimal ones. For highly complex, exploratory, or noisy multi-modal datasets (e.g., OGBench mixed datasets and D4RL antmaze) where Q-value estimation is noisy, a moderate threshold ($\tau \in \{0.6, 0.7\}$) is strictly required to balance suboptimal mode suppression with sufficient behavioral anchoring ($1 - \tau \in \{0.3, 0.4\}$), preventing policy collapse from Q-overestimation. Conversely, for datasets with cleaner distributions or expert-level demonstrations (e.g., D4RL expert or high-quality Minari tasks) with reliable critic gradients, a higher threshold ($\tau \in \{0.8, 0.9\}$) is preferred to maximize peak mode extraction efficiency.
- **VRD-R (Latent Reorganization):** The geometric isolation of high-value generative trajectories in the latent space is governed by the margin parameter σ, searched over a logarithmic grid $\sigma \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. In tasks exhibiting severe Out-of-Distribution (OOD) mode averaging risks, such as OGBench humanoid control, a relatively larger margin ($\sigma = 10^{-2}$) is enforced to guarantee a wider geometric decoupling. For simpler, structured tasks (e.g., D4RL locomotion expert datasets), a strictly smaller margin ($\sigma = 10^{-5}$) prevents over-regularization and maintains the natural topological structure of the generative manifold.

A detailed breakdown of the exact parameters for VRD-Q and VRD-R across specific benchmark tasks is provided in Table 7.

Table 6: General network architectures and optimization hyperparameters shared across all benchmarks.

Category	Hyperparameter	Value
Architecture	Actor / Critic Hidden Layers	$[256, 256, 256]$
	Activation Function	ReLU / SiLU
	Time Embedding Dimension	64
Optimization	Actor / Critic Learning Rate	3×10^{-4}
	Batch Size	256
	Discount Factor (γ)	0.99
	Target Soft-Update Rate (ρ)	0.005
Flow Matching	Integration Steps (NFE) / Solver	10 / Euler
	Noise Schedule	Linear

C.7 Network structures

The architectural framework incorporates the latent variable z across different policy paradigms to enable stochasticity and generative modeling:

One-Step Policy (Actor): Following FQL [21], the latent variable $z \sim \mathcal{N}(0, I)$ (or $z' \sim \mathcal{P}_{\text{test}}$) is concatenated directly with the state observation s as an explicit input feature to the primary MLP.

Table 7: Task-specific hyperparameter selections for VRD-Q (τ) and VRD-R (σ). Highly multi-modal tasks with sparse expert data (e.g., OGBench and AntMaze) utilize a smaller τ and a larger σ (e.g., 10^{-2}), whereas near-expert, structured datasets utilize a larger τ and a strictly smaller σ (e.g., 10^{-5}).

Benchmark	Specific Datasets / Environments	VRD-Q Threshold (τ)	VRD-R Margin (σ)
OGBench	`humanoid-medium`, `ant-large`, `cube-single`, `puzzle`	0.7	10^{-3}
	`humanoid-large`, `ant-giant`, `cube-double`, `scene`	0.6	10^{-2}
D4RL	`antmaze-umaze/medium/large`	$0.6 \sim 0.7$	$10^{-2} \sim 10^{-3}$
	`halfcheetah/hopper/walker2d-medium/medium-replay`	$0.7 \sim 0.8$	10^{-4}
	`halfcheetah/hopper/walker2d-expert/medium-expert`	$0.8 \sim 0.9$	10^{-5}
Minari	Mixed / Sub-optimal datasets (e.g., `pointmaze-medium-v1`)	$0.7 \sim 0.8$	10^{-4}
	Expert datasets (e.g., `pointmaze-expert-v1`, `door-human`)	$0.8 \sim 0.9$	10^{-5}

This allows the feed-forward network $\pi(s, z)$ to nonlinearly warp the latent prior into complex, multi-modal action distributions in a single forward pass without iterative ODE integration.

Flow Policy (ActorVectorField): The latent variable z_t is treated as an explicit input feature alongside observations s and time t. Because the network must model the local velocity of a continuous transformation, z_t represents the intermediate state of the action as it "flows" from the base distribution to the target manifold. The MLP computes the vector field $\dot{z}_t = v_\theta(s, z_t, t)$ based on this concatenated representation.

Critic Network (Value): The critic acts as a deterministic evaluator for a given state-action pair (s, a). While it does not directly process the latent variable z, it estimates the expected return of the final action a (the endpoint of the flow) to guide the policy's optimization.

Table 8: Refined Architectural Specifications Including Latent Variable Roles

Specification	One-Step Policy (`Actor`)	Flow Policy (`ActorVectorField`)	Critic (`Value`)
Primary Inputs	$s \oplus z$ (Obs, Latent z)	$s \oplus z_t \oplus t$ (Obs, Latent z, Time t)	$s \oplus a$ (Obs, Actions)
Latent Variable (z) Role	**Input Feature:** Nonlinearly mapped via MLP to generate one-step action $a = \pi(s, z)$	**Input Feature**: Actively processed to compute flow velocity ($\dot{z}_t$)	**N/A**: Evaluates the final outcome action a
Backbone	MLP (Optional Layer Norm)	Time-conditioned MLP (Sinusoidal Embedding)	Ensemblized MLP ($N \geq 1$ via `vmap`)
Output Type	Deterministic Action a (conditioned on s and z)	Vector Field $v_\theta(s, z_t, t)$	Scalar Value $Q(s, a)$ or $V(s)$

C.8 Computational Resources

All experiments, including the training of the proposed VRD frameworks and the reproduction of baseline methods, were conducted on a high-performance computing workstation to ensure reproducibility and efficiency. Specifically, the deep neural networks and flow-matching policies were accelerated using NVIDIA GeForce RTX 3090 GPUs (24GB VRAM). The algorithms were implemented in Python utilizing the PyTorch/JAX deep learning frameworks. Furthermore, the environment simulations and data sampling for the D4RL and OGBench benchmarks were processed on a Intel Xeon W-2295 CPU equipped with 128GB of RAM.

D Discussion

While the proposed Value-Rectified Distillation (VRD) framework, including its two instantiations (VRD-Q and VRD-R), successfully addresses the fundamental *Latent Fusion* and Out-of-Distribution (OOD) Mode Averaging problems in flow-based offline RL, there are several limitations and open challenges that warrant further investigation.

1. Reliance on Q-value Accuracy. The core mechanism of VRD inherently relies on the accuracy of the underlying value network to guide the trajectory alignment. Specifically, the quantile weight $\omega_\tau(z)$ in VRD-Q and the geometric reshaping function $\Phi_Q(s, z)$ in VRD-R are directly derived from the estimated Q-values. If the Q-network suffers from severe OOD overestimation—a notoriously common issue in offline RL—the VRD framework might inadvertently align the generative flow toward falsely optimistic, low-density regions. Although our empirical results demonstrate robustness, future work could explore integrating more rigorous conservative value estimation methods (e.g., CQL-style penalties or uncertainty-aware ensembles) directly into the value-rectification functional to mitigate this risk.

2. Restriction to Continuous Action Spaces. The theoretical foundation of VRD (as proven in Theorem 1 and 2) and the Flow Matching parameterization are built strictly upon continuous vector fields and continuous topological mappings. Consequently, the current framework is restricted to continuous control tasks. Extending this value-guided flow rectification paradigm to discrete action spaces (e.g., Atari games or discrete routing problems) is non-trivial. Future research could explore incorporating Categorical Flow Matching or Discrete Diffusion processes, adapting the latent distribution alignment functional to discrete categorical probabilities.

3. Potential Diversity Collapse in VRD-R for Multi-Modal Experts. By design, the geometric rectification in VRD-R utilizes a centripetal gravitational field, as Eq.(12), that maps all high-value behaviors towards the origin of the latent space ($\|z\| \approx 0$). While this strictly isolates optimal modes from sub-optimal noise, it may lead to an unintended *diversity collapse* if the underlying optimal behavior itself is inherently multi-modal (i.e., there are multiple distinct but equally optimal strategies). By forcing all optimal trajectories into a single highly compact hypersphere, VRD-R might destroy the topological diversity of expert modes. VRD-Q avoids this by employing an *in-place* statistical filtering, which explains why VRD-Q generally exhibits broader stability across diverse environments. Future iterations of VRD-R could explore multi-centric gravitational fields (e.g., using Gaussian Mixture priors) to preserve expert multi-modality.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The "abstract", "introduction" and "preliminaries and motivation" sections clearly state the problem of latent fusion, and the methodologies to solve this effectively. Theoretical and experimental evidences are provided to support our claims.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated "Limitations" paragraph in the Conclusion section and Appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The main paper explicitly states the required assumptions.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Detailed experimental settings are referenced in the main text. These are extensively expanded upon in Appendices C, providing the necessary hyperparameters, optimizers, and architectural configurations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code were supplied to reviewers as an anonymous zip archive, and full public links will be added in the camera-ready version.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies baseline architectures, optimization protocols, hardware details, and hyperparameters comprehensively in Experiment section and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Our paper reports appropriate information about statistical significance of the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix C details the specific compute hardware used (NVIDIA RTX-3090 GPUs).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research involves fundamental algorithmic developments evaluated on standard, publicly available datasets (e.g., D4RL, Minari, OGBench) and does not violate any ethical guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper focuses on foundational theoretical improvements to flow-based offline RL algorithms. It does not present direct, foreseeable negative societal impacts or obvious ties to malicious applications.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The models and evaluation benchmarks used do not fall under the category of high-risk for dual-use or misuse. No scraped or sensitive data is released.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Existing datasets, platforms, and baseline models are properly credited with citations in the text and the references section.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release new datasets or proprietary assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This research does not involve any human subjects or crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: There is no research involving human subjects, making IRB approval unnecessary.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Large language models are not part of the paper's core methods.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.