SDRB: A SYNTHETIC BENCHMARK FOR DISTRIBUTIONAL REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Distributional reinforcement learning (DRL) models complete return distributions, yet standard benchmarks primarily evaluate control performance rather than the fidelity of the learned distribution. We introduce the Synthetic Distributional Reinforcement Learning Benchmark (SDRB), a controlled testbed for directly evaluating return-distribution reconstruction against exact or high-precision ground truth. SDRB isolates six statistical regimes: Gaussian, mixture, heavy-tailed, sparse, distribution drift, and zero-shot out-of-distribution evaluation. We demonstrate the benchmark through a 540-run factorial study crossing three quantile representations (QR-DQN, IQN, and FQF) with three learning objectives (Quantile Huber, MMD, and Sinkhorn divergence). The experiments reveal substantial representation–objective interactions, reconstruction errors in tail and point-mass structure that are not fully captured by global metrics, failure to recover under persistent stale replay after distribution drift, and mean-preserving OOD shifts that substantially increase distributional error while leaving expected-value error essentially unchanged. These results demonstrate how controlled reconstruction evaluation can expose behavior obscured by aggregate control or expected-value metrics. SDRB complements complex control benchmarks by treating return-distribution reconstruction as a distinct evaluation axis.

1 INTRODUCTION

Distributional reinforcement learning (DRL) models the distribution of cumulative returns rather than only its expectation (Bellemare et al., 2017a). This richer prediction can represent uncertainty, asymmetry, multimodality, and tail behavior that a scalar value cannot capture. Modern DRL methods differ both in how they represent this distribution—for example, through fixed quantiles in QR-DQN (Dabney et al., 2018b), sampled quantile fractions in IQN (Dabney et al., 2018a), or learned fractions in FQF (Yang et al., 2019)—and in the objectives used to learn it.

Evaluating the fidelity of these learned distributions remains difficult. Standard RL benchmarks primarily measure whether an agent achieves high return, but successful control does not require every aspect of the return distribution to be reconstructed accurately. A learned distribution can also have an accurate expectation while misrepresenting its dispersion, tails, or other structure. Moreover, representation and learning objective can interact in ways that are difficult to isolate when algorithms are evaluated only as complete systems in complex environments. Prior work has studied representation and probability-metric choice, return-distribution estimation, and the accuracy of learned distributional predictions (Rowland et al., 2019; Théate et al., 2023; Jullien et al., 2025; Prasad, 2026). What is useful in addition is a controlled experimental setting in which the statistical properties of the target return distribution can themselves be varied while reconstruction is measured against known ground truth.

We introduce the *Synthetic Distributional Reinforcement Learning Benchmark* (SDRB), a diagnostic benchmark designed for this purpose. SDRB uses deliberately simple finite-horizon MDPs with deterministic transitions, while actions determine reward distributions. This removes much of the exploration and transition complexity of conventional control benchmarks and makes the return distribution the primary object of evaluation. Six regimes—Gaussian, Mixture, Heavy-Tail, Sparse, Drift, and out-of-distribution (OOD) evaluation—isolate different forms of statistical complexity, with exact or high-precision reference return distributions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We demonstrate SDRB through a factorial study crossing three quantile representations (QR-DQN, IQN, and FQF) with three learning objectives (Quantile Huber, multiscale MMD, and Sinkhorn divergence) across six regimes and ten random seeds, for $3 \times 3 \times 6 \times 10 = 540$ runs. Wasserstein-1 distance provides a representation-agnostic global reconstruction metric, while regime-specific diagnostics measure properties such as tail and point-mass reconstruction, adaptation after drift, and zero-shot OOD reconstruction.

The study illustrates several kinds of behavior that controlled reconstruction evaluation can expose. First, reconstruction accuracy depends strongly on the representation–objective combination: MMD with QR-DQN achieves the lowest W_1 across all four stationary regimes, but the objective ordering changes under FQF. Second, global or expected-value accuracy can conceal structural errors. Quantile Huber exhibits pronounced underdispersion for QR-DQN and IQN, Heavy-Tail and Sparse reveal errors in known distributional structure, and a mean-preserving OOD scale shift substantially increases reconstruction error without changing expected-value error. Third, under the Drift protocol, all nine configurations fail to recover when obsolete pre-drift samples persist in replay, showing that changing representation or objective alone is insufficient under this condition.

SDRB is intended to complement rather than replace complex control benchmarks. Its purpose is not to identify a universally superior DRL algorithm, but to provide a white-box experimental setting in which reconstruction behavior can be isolated and diagnosed. Our results motivate treating *return-distribution reconstruction* as an evaluation axis distinct from expected-value accuracy and control performance.

2 Related Work

Distributional reinforcement learning. Distributional reinforcement learning models the distribution of cumulative returns rather than only its expectation. Bellemare et al. (2017a) established the modern distributional RL framework and introduced the categorical C51 algorithm. Subsequent methods developed increasingly flexible quantile representations: QR-DQN learns return values at fixed quantile fractions (Dabney et al., 2018b), IQN parameterizes the quantile function and samples fractions during learning (Dabney et al., 2018a), and FQF additionally learns the fraction locations themselves (Yang et al., 2019). Theoretical analyses have studied the approximation properties of categorical and quantile-based distributional methods (Rowland et al., 2018; 2019), while comparisons between expected and distributional RL have shown that improved control performance need not follow directly from more accurate distributional modelling (Lyle et al., 2019). SDRB focuses on a complementary question: how accurately do such representations reconstruct the return distribution itself when the target distribution is known?

Distribution metrics and learning objectives. The geometry used to compare return distributions is a second major design choice in distributional RL. Wasserstein distances play a central role in the original distributional formulation and quantile-based methods (Bellemare et al., 2017a; Dabney et al., 2018b), while the Cramér distance has been studied for its favorable optimization and sample-gradient properties (Bellemare et al., 2017b). Maximum mean discrepancy (MMD) provides a kernel-based discrepancy between probability measures (Gretton et al., 2012) and has been used directly as a distributional RL objective through moment matching (Nguyen-Tang et al., 2021). More recently, SinkhornDRL introduced entropically regularized Wasserstein loss as a distributional RL objective and emphasized the dependence between distribution divergence and representation (Sun et al., 2024). Closest to our factorial motivation, Théate et al. (2023) explicitly decompose a distributional RL algorithm into its representation and probability metric and use a shared monotonic-network framework to compare several metrics. SDRB does not propose a new discrepancy or representation; instead, it provides a controlled benchmark in which established representations and objectives can be crossed factorially and evaluated against known return distributions under multiple statistical regimes.

Evaluation and benchmarking of return distributions. Estimating complete return distributions predates modern deep distributional RL. Morimura et al. (2010b;a) studied parametric and particle-based return-distribution estimation, and Rowland et al. (2019) later provided a unifying view of distributional RL in terms of statistics and distribution imputation. Recent work has studied distributional policy evaluation more directly. Wang et al. (2023) derive tractable return distributions and

approximation errors in distributional LQR, while Zhang et al. (2025) study statistical estimation and inference for complete policy-conditioned return distributions under Wasserstein, Kolmogorov, and total-variation metrics. Gerstenberg et al. (2024) analyze numerical policy-evaluation methods for broad reward mechanisms, including potentially heavy-tailed rewards. Recent empirical work also highlights that learned distributional predictions can misrepresent important distributional structure: Jullien et al. (2025) observe collapse toward the mean under Huber-based quantile learning, and Prasad (2026) audits learned risk claims against Monte Carlo ground truth in standard benchmark environments. SDRB complements these approaches by controlling the statistical structure of the return law itself, allowing reconstruction phenomena to be isolated across representations, learning objectives, and distributional regimes.

Controlled and diagnostic RL benchmarks. Reinforcement-learning benchmarks range from broad performance-oriented platforms to deliberately controlled diagnostic environments. The Atari Learning Environment provides a standardized suite for evaluating general agents (Bellemare et al., 2013), while Procgen emphasizes generalization beyond the finite set of environments observed during training (Cobbe et al., 2020). In contrast, bsuite uses small, carefully designed experiments to isolate specific behavioral capabilities of RL agents (Osband et al., 2020), and MDP Playground provides configurable environments in which sources of RL difficulty can be varied independently (Rajan et al., 2023). More recently, Synthetic Monitoring Environments emphasize white-box evaluation through configurable tasks with known optimal policies and controlled within- and out-of-distribution analysis (Pleiss et al., 2026). SDRB adopts this diagnostic philosophy but targets a different object of study: the fidelity with which distributional RL methods reconstruct return distributions. By combining controlled reward laws with exact or high-precision return ground truth, SDRB is intended to complement complex control benchmarks with a testbed for diagnosing distributional reconstruction behavior.

3 METHODOLOGY

3.1 SYNTHETIC DISTRIBUTIONAL REINFORCEMENT LEARNING BENCHMARK

We introduce the *Synthetic Distributional Reinforcement Learning Benchmark* (SDRB), a controlled benchmark for evaluating how accurately distributional reinforcement learning (DRL) methods reconstruct return distributions. In conventional reinforcement-learning benchmarks, the learned return distribution is typically evaluated indirectly through policy performance. This makes it difficult to distinguish improvements in distributional representation from improvements in control, exploration, or representation learning. SDRB instead constructs small Markov decision processes for which the statistical structure of the return distribution is explicitly controlled and ground-truth return distributions can be computed exactly or to high numerical precision.

Our base environment is a deterministic finite-horizon chain with horizon $H = 10$,

$$s_0 \to s_1 \to \cdots \to s_H,$$

where s_H is terminal. Thus, a trajectory beginning at s_0 contains H reward-emitting transitions. At each nonterminal state, the agent selects between two actions. Actions do not affect the transition dynamics; instead, they determine the reward distribution. This design deliberately separates return-distribution learning from exploration and stochastic transition branching.

For a fixed policy, the return from state s_t is

$$G_t = \sum_{k=t}^{H-1} \gamma^{k-t} R_k,$$

where $\gamma = 0.99$. Because the successor-state sequence is deterministic, return stochasticity in SDRB arises from the controlled reward-emission processes accumulated along a non-branching trajectory. SDRB therefore isolates temporal distributional bootstrapping under reward-channel stochasticity; it does not attempt to model multimodality or uncertainty introduced by stochastic branching among successor states.

We evaluate a fixed policy selecting action 1. The resulting task is controlled distributional policy evaluation: the learner must reconstruct the policy-conditioned return distribution rather than

Table 1: SDRB regimes and the distributional properties they isolate.

Regime	Controlled property	Primary diagnostic
Gaussian	Smooth unimodal returns	Location / dispersion
Mixture	Non-Gaussian structure	Global reconstruction
Heavy Tail	Rare extreme outcomes	Tail reconstruction
Sparse	Discrete–continuous structure	Zero-mass structure
Drift	Abrupt nonstationarity	Adaptation / recovery
OOD	Unseen reward laws	Zero-shot reconstruction

simultaneously solving a difficult control problem. Ground-truth return distributions are computed for each state and action. Exact distributions are used where analytically tractable, and fixed-seed high-precision Monte Carlo approximations are used otherwise.

3.2 Distributional Regimes

SDRB contains six regimes that isolate distinct statistical properties of return distributions. Table 1 summarizes their roles; complete reward-law specifications and ground-truth procedures are provided in Appendix A. The numerical reward-process parameters used in the primary factorial study are reported in Appendix A, Table 4.

Gaussian rewards provide an analytically tractable calibration case, while Gaussian mixtures introduce structured non-Gaussian returns. Heavy-Tail rewards use Gaussian contamination,

$$R \sim (1-p)\mathcal{N}(\mu_b, \sigma_b^2) + p\mathcal{N}(\mu_t, \sigma_t^2), \tag{1}$$

to introduce rare extreme outcomes. Sparse rewards follow a zero-inflated Gaussian law,

$$R = \begin{cases} 0, & \text{with probability } 1-p, \\ X, \quad X \sim \mathcal{N}(\mu, \sigma^2), & \text{with probability } p, \end{cases} \tag{2}$$

producing a mixed discrete–continuous return distribution.

Drift introduces an abrupt change in the reward law at a fixed environment step while leaving the transition dynamics unchanged. OOD is an evaluation protocol rather than a separate training world: models trained on the Gaussian regime are frozen and evaluated without adaptation under unseen location, scale, and combined shifts. In particular, the scale shift preserves the reward mean while changing its variance, allowing distributional error to be isolated from expected-value error.

3.3 Distributional Representations

We compare three quantile-based representations with $N = 51$ atoms. We intentionally restrict the present factorial representation axis to the quantile family so that objective changes can be studied across methods that share an unconstrained-support quantile interpretation. Categorical probability-mass representations such as C51 provide a complementary representation geometry and are outside the scope of the present factorial study. **QR-DQN** learns return values at fixed quantile fractions. **IQN** parameterizes the quantile function and samples 51 fractions during training; evaluation uses a deterministic 51-point midpoint grid. **FQF** additionally learns quantile-fraction locations through a fraction-proposal network, with interval widths

$$w_i = \tau_{i+1} - \tau_i \tag{3}$$

defining the probability masses of its empirical return distribution.

FQF's fraction and value networks are optimized separately using the standard fraction objective. We calibrated the fraction learning rate on one Gaussian diagnostic before the full study and then fixed it at $\eta_{\text{fraction}} = 10^{-5}$; Appendix C reports the calibration and fraction diagnostics.

3.4 Distributional Learning Objectives

We compare three objectives that represent distinct distributional geometries.

Quantile Huber Regression. Our conventional DRL baseline minimizes the Quantile Huber loss between predicted and Bellman-target return atoms.

Maximum Mean Discrepancy. We use biased empirical squared MMD with a multi-scale RBF kernel. A detached target median pairwise distance defines the base bandwidth, with kernel scales

$$\{0.25,\ 0.5,\ 1,\ 2,\ 4\}. \tag{4}$$

MMD is permutation invariant and directly compares empirical return measures.

Sinkhorn Divergence. We used the debiased entropic Sinkhorn divergence

$$S_\varepsilon(P,Q) = \mathrm{OT} * \varepsilon(P,Q) - \frac{1}{2}\mathrm{OT} * \varepsilon(P,P) - \frac{1}{2}\mathrm{OT} * \varepsilon(Q,Q), \tag{5}$$

with ground cost $|x-y|$ and 50 Sinkhorn iterations. The regularization coefficient is computed from the detached target geometry as

$$\varepsilon = \max\left(10^{-5}, 0.003 \operatorname*{median}_{i\neq j} |y_i - y_j|\right).$$

The multiplicative factor 0.003 was selected in the calibration described in Appendix C.3 and then frozen for the primary factorial study.

QR-DQN and IQN induce equal-weight empirical measures, whereas FQF uses its learned interval masses. We therefore use weighted MMD and Sinkhorn for FQF, with masses detached during value optimization. Calibration and gradient-isolation tests are reported in Appendices C and D. These three objectives are intended to sample distinct learning geometries rather than constitute an exhaustive survey of distributional discrepancies. Other objectives, including Cramér-based losses, provide natural extensions of the benchmark.

3.5 Evaluation Metrics

Our primary representation-agnostic reconstruction metric is the one-dimensional Wasserstein-1 distance, reported in return units. Unless otherwise stated, stationary comparisons use the final checkpoint at s_0, the state with the longest remaining Bellman depth.

Secondary generic metrics include CDF error, support coverage, mean error, and the predicted-to-true standard-deviation ratio. We treat quantile MAE and quantile crossing as representation diagnostics rather than independent reconstruction metrics.

Regime-specific diagnostics target known structural properties of each benchmark. Heavy Tail evaluates lower- and upper-tail quantile error and exceedance probability; Sparse evaluates predicted zero mass and error over the analytically known zero-return plateau; Drift evaluates pre-change error, recovery, adaptation lag, and post-drift error area. We use the 0.01 and 0.99 quantiles for tail diagnostics, a $|\hat{z}| \leq 0.05$ tolerance for predicted zero mass, and define Drift recovery relative to the final five pre-drift checkpoints. Complete operational definitions are given in Appendix A.5.

3.6 Experimental Protocol

We evaluate 3 representations $\times$ 3 objectives $\times$ 6 regimes $\times$ 10 seeds $= 540$ runs, using seeds 123–132. Each run uses $H = 10$, $\gamma = 0.99$, 1,000 episodes, eight parallel environments, and 51 return locations. All objective comparisons share the same replay, optimization, target-update, evaluation, and training protocols. Replay stores raw one-step transitions and distributional Bellman targets are constructed during learner updates. Complete architecture and optimization hyperparameters are reported in Appendix A.4.

4 Results

4.1 Reconstruction Accuracy Depends on Both Objective and Representation

Table 2 reports final state-s_0 Wasserstein-1 reconstruction error across the four stationary SDRB regimes. No objective produces the same relative behavior across all three representations. For

Table 2: Final state-0 Wasserstein-1 reconstruction error (mean ± standard deviation over 10 seeds; lower is better). Bold denotes the best mean within each benchmark.

Agent	Objective	Gaussian	Mixture	Heavy Tail	Sparse
QR-DQN	Quantile Huber	0.801 ± 0.122	1.042 ± 0.069	1.160 ± 0.161	1.129 ± 0.032
QR-DQN	MMD	$\mathbf{0.242 \pm 0.141}$	$\mathbf{0.431 \pm 0.245}$	$\mathbf{0.588 \pm 0.157}$	$\mathbf{0.316 \pm 0.098}$
QR-DQN	Sinkhorn	0.595 ± 0.245	0.798 ± 0.241	1.654 ± 0.101	2.743 ± 0.395
IQN	Quantile Huber	0.786 ± 0.080	1.065 ± 0.090	1.069 ± 0.168	1.110 ± 0.034
IQN	MMD	0.374 ± 0.180	0.667 ± 0.278	0.960 ± 0.264	0.366 ± 0.096
IQN	Sinkhorn	0.583 ± 0.398	0.733 ± 0.513	1.493 ± 0.109	3.033 ± 0.643
FQF	Quantile Huber	0.553 ± 0.231	0.903 ± 0.269	1.070 ± 0.326	0.865 ± 0.155
FQF	MMD	1.883 ± 0.529	2.128 ± 1.009	1.233 ± 0.324	0.541 ± 0.141
FQF	Sinkhorn	1.270 ± 0.427	1.030 ± 0.531	1.014 ± 0.333	1.178 ± 0.586

QR-DQN, MMD achieves the lowest W_1 in Gaussian, Mixture, Heavy-Tail, and Sparse, with errors of 0.242, 0.431, 0.588, and 0.316, respectively. The same objective improves IQN substantially on Gaussian and Sparse, although its advantage is smaller on Mixture and Heavy-Tail. Under FQF, however, the ordering changes. Quantile Huber achieves the lowest mean error on Gaussian and Mixture, while MMD achieves the lowest mean on Sparse. On Heavy Tail, Quantile Huber and Sinkhorn are comparable within the observed seed variability, with Sinkhorn having the numerically lowest mean (1.014 versus 1.070).

This dependence on the representation–objective combination is consistent with prior work treating the representation and probability discrepancy as distinct but interacting design choices in distributional RL (Théate et al., 2023; Sun et al., 2024). SDRB makes this interaction directly measurable under a common environment and evaluation protocol: across the controlled factorial comparison, changing the representation changes the relative ordering of the learning objectives.

The Gaussian regime also reveals qualitatively different reconstruction errors that are not apparent from W_1 alone. Quantile Huber produces predicted-to-true standard-deviation ratios of approximately 0.405 for QR-DQN and 0.408 for IQN, indicating substantial underdispersion. MMD yields ratios much closer to one, approximately 0.979 and 0.938, respectively. Under FQF, the pattern changes again: Quantile Huber is slightly overdispersed, whereas MMD is substantially overdispersed. Recent work has similarly reported collapse toward the mean under Huber-based quantile learning (Jullien et al., 2025). Here, known return distributions allow SDRB to quantify such dispersion errors directly and to show that the pattern depends on the representation with which the objective is paired.

Probability-location controls. Because IQN samples quantile fractions during training while FQF learns their locations adaptively, we tested whether probability-location selection could account for the observed representation–objective interaction. We repeated the four stationary regimes using a deterministic midpoint grid for IQN and fixed uniform fractions for FQF, while otherwise preserving the corresponding architectures and training protocols. The interventions materially change reconstruction error but do not produce a consistent improvement or degradation. Fixing FQF fractions worsens W_1 under Quantile Huber in all four regimes, improves it under MMD in all four, and has regime-dependent effects under Sinkhorn. Fixed-grid IQN likewise exhibits objective- and regime-dependent changes. Probability-location selection therefore contributes to the observed interaction but does not explain it away; its effect depends on the distributional objective and target return law. Appendix B.6 reports the complete control matrix.

4.2 Expected-Value Accuracy Can Conceal OOD Distribution Error

The OOD protocol provides a controlled demonstration of a limitation of expected-value-only evaluation. Agents are trained in the Gaussian in-distribution environment, frozen, and then evaluated without adaptation under unseen reward laws. In the Scale condition, the target variance changes while the target expected value is preserved exactly. Because the predictor is also frozen, its expected-value prediction is unchanged; expected-value error is therefore invariant to this intervention by construction.

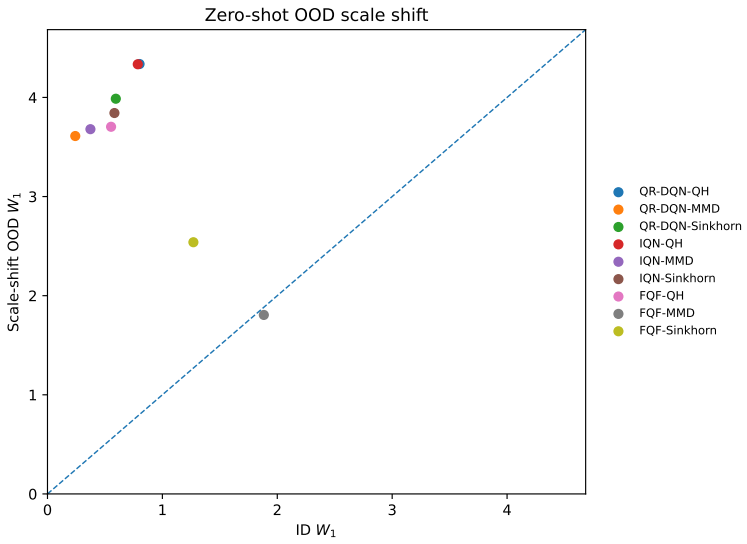

Figure 1: Zero-shot OOD reconstruction under a mean-preserving scale shift. Expected-value error is invariant to this intervention by construction for the frozen predictor, while Wasserstein reconstruction error increases for most configurations.

Distributional reconstruction error is not invariant. For QR-DQN with Quantile Huber, state-s_0 W_1 increases from 0.801 in distribution to 4.337 under the Scale intervention, and IQN increases from 0.786 to 4.335. Most MMD and Sinkhorn configurations similarly exhibit substantial distributional degradation. The experiment should therefore be interpreted as a controlled demonstration that an expectation-based metric is mathematically blind to a mean-preserving change in return dispersion, rather than as evidence that the learned networks independently preserve expected value under OOD shift.

This separation complements prior work distinguishing the benefits of distributional modelling from expected-value or control performance (Lyle et al., 2019) and recent efforts to audit whether learned distributional predictions accurately reflect return uncertainty (Prasad, 2026). SDRB provides a controlled version of this diagnostic: because the OOD shift is constructed to preserve the mean while changing the scale, the resulting reconstruction error can be attributed directly to distributional structure that expected-value evaluation cannot observe.

FQF-MMD provides an instructive exception. Its W_1 decreases slightly from 1.883 in-distribution to 1.806 under the scale shift. This should not be interpreted as adaptation or OOD generalization: the network is frozen, and its in-distribution prediction is already substantially overdispersed. The broader OOD target therefore happens to align more closely with the learned distribution.

4.3 Global Reconstruction Error Does Not Fully Characterize Structural Accuracy

The Heavy-Tail and Sparse regimes illustrate why a global probability metric should be supplemented by diagnostics matched to known statistical structure. Table 3 reports representative regime-specific errors alongside global reconstruction performance.

In Heavy Tail, improved global reconstruction can coincide with worse reconstruction of a specific rare-event feature. For QR-DQN, MMD reduces state-s_0 W_1 from 1.160 under Quantile Huber to 0.588, yet its upper-tail quantile error increases from 4.202 to 4.475. IQN shows the same qualitative tradeoff: MMD modestly improves global W_1 while increasing upper-tail error relative to Quantile Huber. Across all three representations, Quantile Huber produces the lowest mean upper-tail error in Table 3. Thus, optimizing the bulk of the learned distribution more accurately according to a global metric need not improve a localized tail property. This is precisely the type of structural tradeoff that is hidden when evaluation is reduced to a single aggregate reconstruction score.

Table 3: Specialized SDRB diagnostics at state 0. Lower is better except Drift recovered, where one indicates successful recovery.

Agent	Objective	Heavy-tail upper-tail error	Sparse zero-plateau MAE	Drift final W_1	Drift recovered
QR-DQN	Quantile Huber	4.202 ± 0.521	2.662 ± 0.151	18.859 ± 0.735	0.000 ± 0.000
QR-DQN	MMD	4.475 ± 0.585	3.680 ± 0.516	19.172 ± 0.495	0.000 ± 0.000
QR-DQN	Sinkhorn	9.475 ± 0.694	1.119 ± 0.559	20.631 ± 0.865	0.000 ± 0.000
IQN	Quantile Huber	3.161 ± 0.645	2.659 ± 0.120	18.967 ± 0.587	0.000 ± 0.000
IQN	MMD	4.698 ± 1.155	3.177 ± 1.341	19.192 ± 0.447	0.000 ± 0.000
IQN	Sinkhorn	9.386 ± 0.959	0.868 ± 0.825	20.267 ± 0.655	0.000 ± 0.000
FQF	Quantile Huber	2.028 ± 1.115	2.458 ± 0.280	19.423 ± 0.629	0.000 ± 0.000
FQF	MMD	4.254 ± 1.937	2.836 ± 1.127	19.331 ± 0.514	0.000 ± 0.000
FQF	Sinkhorn	5.268 ± 3.461	2.842 ± 0.916	20.422 ± 0.973	0.000 ± 0.000

Sparse exposes a different structural limitation. The true return law contains a discrete zero-return component together with a continuous nonzero component. QR-DQN-MMD and IQN-MMD achieve low global W_1 and dispersion close to the ground truth, but neither exactly reconstructs the known zero-return structure. Conversely, configurations with worse global W_1 can behave differently around the zero plateau. These results show that global reconstruction error alone does not determine whether a method has recovered a known structural feature of the target distribution.

Together, these regimes demonstrate the role of SDRB's specialized diagnostics: when the data-generating process is known, evaluation can ask not only how far two distributions are globally, but also which statistical features account for the remaining reconstruction error.

4.4 All Methods Fail to Recover Under Persistent Stale Replay

The Drift regime is a deliberately severe stale-replay stress condition: pre-drift experience remains in a replay buffer whose capacity exceeds the complete training trajectory. All nine representation–objective combinations fail the recovery criterion, with final state-s_0 W_1 between approximately 18.9 and 20.6.

This result should not be interpreted as a general inability of distributional methods to adapt to nonstationarity. Rather, under persistent uniform replay, representation or objective choice alone is insufficient to overcome the mixture of obsolete and current experience. Appendix B.4 reports the complete recovery diagnostics.

4.5 Discussion: Reconstruction Quality Is a Distinct Evaluation Axis

The factorial study demonstrates SDRB's diagnostic use rather than ranking DRL algorithms. QR-DQN with MMD achieves the lowest W_1 across the four stationary regimes, but objective ordering changes across representations and specialized diagnostics expose errors hidden by global W_1.

FQF also reveals a more specific optimization interaction. Under MMD and Sinkhorn, the value network uses a permutation-invariant empirical-measure objective while the fraction network retains the canonical order-dependent FQF surrogate. Although gradient isolation is verified, these objectives do not enforce the ordered quantile surface assumed by the fraction update. The probability-location controls in Appendix B.6 show that fraction adaptation is consequential but is not a sufficient explanation of the broader representation–objective interaction. Disabling FQF fraction learning improves reconstruction under MMD but degrades it under Quantile Huber, while Sinkhorn exhibits regime-dependent effects. Deterministic IQN fractions also produce objective- and regime-dependent changes. Thus, adaptive probability allocation is neither uniformly beneficial nor uniformly harmful; its effect depends on the learning objective and target-distribution geometry.

Across the experiments, controlled ground truth exposes representation–objective interactions, dispersion and structural errors, persistent stale-replay failure, and OOD distributional degradation invisible to expected-value evaluation. Related work has approached such questions through algorithm design, statistical analysis, and auditing (Rowland et al., 2019; Théate et al., 2023; Jullien et al., 2025; Prasad, 2026); SDRB complements these approaches by controlling the statistical

source of reconstruction difficulty directly. We therefore treat return-distribution reconstruction as an evaluation axis distinct from expected-value accuracy and control performance.

5 Conclusion and Limitations

We introduced SDRB, a controlled benchmark for evaluating return-distribution reconstruction against known ground truth. A 540-run factorial study demonstrates how SDRB exposes representation–objective interactions, structural errors hidden by global metrics, failure under persistent stale replay, and OOD distribution error hidden by expected-value accuracy. These findings motivate treating return-distribution reconstruction as an evaluation axis distinct from expected-value accuracy and control performance. SDRB complements complex control benchmarks with a white-box setting for diagnosing what distributional RL methods actually reconstruct.

SDRB deliberately trades environmental complexity for experimental control: the present benchmark studies reward-channel stochasticity along deterministic, non-branching trajectories and does not test stochastic transition branching, high-dimensional observations, exploration, or complex control. The factorial study is also restricted to three quantile representations and three objectives; categorical, monotonic, hybrid, and other distributional methods are natural extensions rather than targets of the present claims. The Drift result specifically concerns persistent stale replay and should not be generalized to other replay or adaptation mechanisms. Finally, SDRB uses exact return laws where tractable and fixed high-precision Monte Carlo references otherwise.

Reproducibility Statement

SDRB is designed to support reproducible evaluation through explicitly specified synthetic reward processes and controlled experimental protocols. Appendix A reports the numerical reward-law parameters, oracle-construction procedures, training configuration, and operational metric definitions used for the primary factorial study.

AI Use Statement

Generative AI tools were used during this work as research and writing assistants. They were used to provide feedback on experimental methodology and implementation, assist with software development and debugging and support the interpretation and organization of experimental results. Generative AI tools were also used to help draft, edit, and restructure portions of the manuscript and supporting research artifacts. The author reviewed and validated all AI-assisted code and experimental artifacts used in the reported study through testing and empirical verification. The author determined the research questions, made the final methodological and experimental decisions, executed the reported experiments, evaluated the resulting evidence, and takes responsibility for the final content of the paper and all artifacts produced with the aid of generative AI.

References

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of artificial intelligence research*, 47: 253–279, 2013.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pp. 449–458. Pmlr, 2017a.

Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The cramér distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017b.

Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *International conference on machine learning*, pp. 2048–2056. PMLR, 2020.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pp. 1096–1105. PMLR, 2018a.

Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018b.

Julian Gerstenberg, Ralph Neininger, and Denis Spiegel. On policy evaluation algorithms in distributional reinforcement learning. *arXiv preprint arXiv:2407.14175*, 2024.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13:723–773, 2012.

Sami Jullien, Romain Deffayet, Jean-Michel Renders, Paul Groth, and Maarten de Rijke. Distributional reinforcement learning with dual expectile-quantile regression. In Silvia Chiappa and Sara Magliacane (eds.), *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 1909–1923. PMLR, 21–25 Jul 2025. URL https://proceedings.mlr.press/v286/jullien25a.html.

Clare Lyle, Marc G Bellemare, and Pablo Samuel Castro. A comparative analysis of expected and distributional reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 4504–4511, 2019.

Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 799–806, 2010a.

Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Parametric return density estimation for reinforcement learning. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pp. 368–375, 2010b.

Thanh Nguyen-Tang, Sunil Gupta, and Svetha Venkatesh. Distributional reinforcement learning via moment matching. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 9144–9152, 2021.

Ian Osband, Yotam Doron, Matteo Hessel, John Aslanides, Eren Sezener, Andre Saraiva, Katrina McKinney, Tor Lattimore, Csaba Szepesvari, Satinder Singh, Benjamin Van Roy, Richard Sutton, David Silver, and Hado Van Hasselt. Behaviour suite for reinforcement learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygf-kSYwH.

Leonard S. Pleiss, Carolin Schmidt, and Maximilian Schiffer. Synthetic monitoring environments for reinforcement learning. *Reinforcement Learning Journal*, 7, 2026.

Hari Prasad. Auditing the risk claims of distributional reinforcement learning. *arXiv preprint arXiv:2607.11607*, 2026.

Raghu Rajan, Jessica Lizeth Borja Diaz, Suresh Guttikonda, Fábio Ferreira, André Biedenkapp, Jan Ole von Hartz, and Frank Hutter. Mdp playground: An analysis and debug testbed for reinforcement learning. *Journal of Artificial Intelligence Research*, 2023. doi: 10.1613/JAIR.1.14314.

Mark Rowland, Marc Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 29–37. PMLR, 2018.

Mark Rowland, Robert Dadashi, Saurabh Kumar, Rémi Munos, Marc G Bellemare, and Will Dabney. Statistics and samples in distributional reinforcement learning. In *International Conference on Machine Learning*, pp. 5528–5536. PMLR, 2019.

Ke Sun, Yingnan Zhao, Wulong Liu, Bei Jiang, and Linglong Kong. Distributional reinforcement learning with regularized wasserstein loss. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 63184–63221. Curran Associates, Inc., 2024. doi: 10.52202/079017-2018. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7371ee6a40da2951303ec7ebdb2150ce-Paper-Conference.pdf.

Thibaut Théate, Antoine Wehenkel, Adrien Bolland, Gilles Louppe, and Damien Ernst. Distributional reinforcement learning with unconstrained monotonic neural networks. *Neurocomputing*, 534:199–219, 2023. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.02.049. URL https://www.sciencedirect.com/science/article/pii/S0925231223001996.

Zifan Wang, Yulong Gao, Siyi Wang, Michael M Zavlanos, Alessandro Abate, and Karl Henrik Johansson. Policy evaluation in distributional lqr. In *Learning for dynamics and control conference*, pp. 1245–1256. PMLR, 2023.

Derek Yang, Li Zhao, Zichuan Lin, Tao Qin, Jiang Bian, and Tie-Yan Liu. Fully parameterized quantile function for distributional reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

Liangyu Zhang, Yang Peng, Jiadong Liang, Wenhao Yang, and Zhihua Zhang. Estimation and inference in distributional reinforcement learning. *The Annals of Statistics*, 53(5):1987–2011, 2025.

A Additional Benchmark Details

This appendix provides additional details on the construction and evaluation protocol of the Synthetic Distributional Reinforcement Learning Benchmark (SDRB).

A.1 Finite-Horizon Chain

The base MDP is a deterministic chain with horizon $H = 10$,

$$s_0 \to s_1 \to \cdots \to s_H \tag{6}$$

At each nonterminal state, the available actions affect only the reward distribution and do not affect the subsequent state. The transition structure therefore remains fixed across all benchmark regimes.

This design makes it possible to isolate return-distribution estimation from transition uncertainty, exploration difficulty, and high-dimensional representation learning. Under a fixed policy, the return distribution at state s_t is determined entirely by the reward laws encountered along the remaining deterministic trajectory.

All primary experiments use discount factor $\gamma = 0.99$.

A.2 Ground-Truth Return Distributions

Ground-truth return distributions are generated independently of the learned distributional representations. Where the finite-horizon return law is analytically tractable, SDRB uses exact calculations. Gaussian return means and variances are therefore computed analytically.

For regimes requiring numerical reference distributions, oracle generation is performed once using a fixed oracle random seed and the resulting targets are reused across learner seeds and objectives. Heavy-Tail and Sparse oracle targets use 300,000 Monte Carlo return samples per configured reference distribution. Empirical quantiles are obtained from the resulting cached return samples. Fixing the oracle realization prevents target-generation noise from varying across the factorial learner comparison; variability reported in the experiments therefore reflects learner seeds rather than independent Monte Carlo targets.

Ground-truth distributions are available at the state–action level. The primary stationary results in the main paper report the final checkpoint at s_0, which corresponds to the largest remaining Bellman depth.

A.3 Benchmark Regimes

Table 4: Numerical reward-process parameters used in the primary SDRB experiments. All primary policy-evaluation results use action 1. Action 0 is reported because state–action ground truth is available for both actions.

Regime	Action 0	Action 1
Gaussian	$\mathcal{N}(0, 1^2)$	$\mathcal{N}(2, 0.5^2)$
Mixture	$0.5\mathcal{N}(-1, 0.35^2) + 0.5\mathcal{N}(1, 0.35^2)$	$0.5\mathcal{N}(1.25, 0.35^2) + 0.5\mathcal{N}(2.75, 0.35^2)$
Heavy Tail	$0.99\mathcal{N}(0, 1^2) + 0.01\mathcal{N}(0, 10^2)$	$0.98\mathcal{N}(1, 0.5^2) + 0.02\mathcal{N}(8, 1^2)$
Sparse	$B\mathcal{N}(1, 0.25^2),\ B \sim \text{Bernoulli}(0.1)$	$B\mathcal{N}(2, 0.25^2),\ B \sim \text{Bernoulli}(0.2)$
Drift, pre-change	$\mathcal{N}(0, 1^2)$	$\mathcal{N}(2, 0.5^2)$
Drift, post-change	$\mathcal{N}(0, 1^2)$	$\mathcal{N}(-1, 0.5^2)$

Gaussian. The Gaussian regime provides the simplest smooth unimodal target and serves as the primary diagnostic environment used before large-scale evaluation. Action 0 uses a Gaussian reward law centered at zero, while the evaluated action uses a shifted lower-variance Gaussian reward law.

Mixture. The Mixture regime introduces non-Gaussian return structure through mixtures of Gaussian reward components. We evaluate the resulting complete return distribution rather than attempting to associate learned atoms with individual mixture components.

Heavy Tail. The Heavy-Tail regime uses Gaussian contamination,

$$R \sim (1-p)\mathcal{N}(\mu_b, \sigma_b^2) + p\mathcal{N}(\mu_t, \sigma_t^2). \tag{7}$$

The evaluated action uses a low-probability shifted tail component. This creates rare but consequential outcomes while preserving a simple generative description.

Sparse Reward. The Sparse regime uses a zero-inflated Gaussian reward process,

$$R = \begin{cases} 0, & \text{with probability } 1-p, \\ X, \quad X \sim \mathcal{N}(\mu, \sigma^2), & \text{with probability } p. \end{cases} \tag{8}$$

The exact zero-return mass for a fixed horizon is determined by the probability that no nonzero reward event occurs across the remaining trajectory. This produces a mixed discrete–continuous return law.

Distribution Drift. The Drift regime introduces an abrupt change after one half of the configured training trajectory. Before the change, action 1 follows $\mathcal{N}(2, 0.5^2)$; afterward it follows $\mathcal{N}(-1, 0.5^2)$. Action 0 remains $\mathcal{N}(0, 1^2)$ throughout. The transition dynamics do not change. Because replay capacity exceeds the complete training trajectory, pre-drift transitions are retained for the remainder of training. The experiment is therefore a deliberately severe persistent stale-replay condition rather than a general test of nonstationary adaptation.

OOD Evaluation. OOD evaluation reuses the Gaussian family as an evaluation protocol rather than introducing a separate training environment. Networks are trained under the in-distribution action-1 law $\mathcal{N}(2, 0.5^2)$ and then frozen. Action 0 remains $\mathcal{N}(0, 1^2)$ in every case. The action-1 evaluation laws are

Case	μ	σ
ID	2	0.5
Location +	4	0.5
Location –	0	0.5
Scale	2	2
Combined	4	2.

The Scale intervention is mean preserving by construction. Consequently, for a frozen predictor, expected-value error is unchanged by this intervention while distributional quantities sensitive to dispersion may change substantially.

A.4 Replay and Training Protocol

The replay buffer stores raw one-step transitions rather than precomputed discounted returns. Distributional Bellman targets are constructed during learner updates.

The same replay capacity, warm-up period, batch size, target-network update schedule, environment horizon, and training budget are used across objective comparisons unless an agent architecture requires an architecture-specific component.

For Heavy Tail, lower- and upper-tail quantile errors are $|\hat{Q}(0.01) - Q(0.01)|$ and $|\hat{Q}(0.99) - Q(0.99)|$, with tail qMAE equal to their mean.

For Sparse, predicted zero mass is the empirical probability assigned to $|\hat{z}| \leq 0.05$. Zero-plateau MAE is the mean absolute quantile error over quantile locations within the analytically known zero-return probability region.

For Drift, let e_j denote state-s_0 W_1 at checkpoint j. The pre-drift baseline $\bar{e}_{\text{pre}}$ is the mean of the final five pre-drift checkpoints and the recovery threshold is $T_{\text{rec}} = 1.25\bar{e}_{\text{pre}}$. Recovery occurs at the first recorded post-drift checkpoint satisfying $e_j \leq T_{\text{rec}}$; no additional persistence window is imposed. Adaptation lag is the number of environment transitions from the drift step to that checkpoint. Post-drift error area is the trapezoidal integral of $\max(e_j - \bar{e}_{\text{pre}}, 0)$ over environment transitions.

Table 5: Common primary training configuration.

Parameter	Value
Horizon H	10
Discount γ	0.99
Episodes	1000
Parallel environments	8
Return locations / quantiles	51
Value learning rate	10^{-3}
Batch size	64
Replay capacity	100,000
Replay warmup	1,000 transitions
Target-update interval	500
Learner updates per vector step	1
Hidden dimension	128
State embedding dimension	32
Gradient-norm limit	10
Quantile-Huber κ	1
Evaluation interval	50 episodes
Seeds	123–132
FQF fraction learning rate	10^{-5}
FQF cosine embedding dimension	64
FQF fraction entropy coefficient	10^{-3}

A.5 Primary and Secondary Metrics

The primary reconstruction metric is the one-dimensional Wasserstein-1 distance. Secondary generic metrics include CDF error, support coverage, mean error, and the predicted-to-true standard-deviation ratio.

Representation-specific diagnostics such as quantile mean absolute error and quantile-crossing rate are not treated as independent generic reconstruction metrics. In particular, crossing rate has a different interpretation for fixed ordered quantile heads and permutation-invariant empirical-distribution objectives.

Heavy-Tail results additionally include tail-quantile and exceedance diagnostics. Sparse results include predicted zero mass, zero-mass error, and error around the zero plateau. Drift results include pre-drift error, peak post-drift error, recovery status, adaptation lag, and post-drift area under the reconstruction-error curve.

All main stationary comparisons use the final checkpoint at s_0.

B Additional Experimental Results

This appendix supplements the main-paper results with additional diagnostics intended to distinguish global reconstruction quality from representation-specific behavior.

B.1 Stationary Reconstruction Beyond Wasserstein Error

The main paper reports Wasserstein-1 error as the primary representation-agnostic metric. Secondary metrics show that similar global errors can correspond to different reconstruction failures. Under quantile Huber regression, QR-DQN and IQN are consistently underdispersed on the stationary Gaussian regime. Their predicted-to-true standard-deviation ratios are approximately 0.405 and 0.408, respectively. In contrast, MMD produces ratios closer to one: approximately 0.979 for QR-DQN and 0.938 for IQN. FQF exhibits stronger objective-dependent variation. Under Quantile Huber regression, its Gaussian standard-deviation ratio is approximately 1.09, whereas weighted MMD produces substantial overdispersion, with a ratio of approximately 2.60. Weighted Sinkhorn also produces overdispersion, although less strongly. These differences reinforce that Wasserstein error should be interpreted together with dispersion and support diagnostics when identifying the mechanism of failure.

Table 6: Gaussian state-0 reconstruction diagnostics (mean $\pm$ standard deviation over 10 seeds). W_1, mean error, and CDF MAE are lower-is-better; a standard-deviation ratio of one indicates correct dispersion.

Agent	Objective	W_1	Mean error	Pred./true std.	CDF MAE
QR-DQN	Quantile Huber	0.801 ± 0.122	0.331 ± 0.318	0.405 ± 0.025	0.033 ± 0.005
QR-DQN	MMD	0.242 ± 0.141	0.216 ± 0.160	0.979 ± 0.047	0.010 ± 0.006
QR-DQN	Sinkhorn	0.595 ± 0.245	0.507 ± 0.310	0.694 ± 0.045	0.025 ± 0.010
IQN	Quantile Huber	0.786 ± 0.080	0.315 ± 0.255	0.408 ± 0.025	0.033 ± 0.003
IQN	MMD	0.374 ± 0.180	0.345 ± 0.201	0.938 ± 0.085	0.016 ± 0.007
IQN	Sinkhorn	0.583 ± 0.398	0.533 ± 0.423	0.812 ± 0.058	0.024 ± 0.016
FQF	Quantile Huber	0.553 ± 0.231	0.378 ± 0.328	1.092 ± 0.123	0.023 ± 0.010
FQF	MMD	1.883 ± 0.529	0.376 ± 0.277	2.596 ± 0.464	0.077 ± 0.021
FQF	Sinkhorn	1.270 ± 0.427	0.572 ± 0.308	1.976 ± 0.344	0.052 ± 0.018

B.2 Heavy-Tail Diagnostics

Heavy-Tail evaluation separates central-distribution accuracy from rare-event reconstruction. For QR-DQN with quantile Huber regression, the final state-s_0 Wasserstein error is approximately 1.16, while the upper-tail quantile error remains substantially larger. QR-DQN with MMD improves global reconstruction to approximately 0.59 but retains nontrivial tail error. Similar behavior appears for IQN. Sinkhorn exhibits particularly weak tail reconstruction for several agent configurations. FQF-Sinkhorn obtains competitive global Heavy-Tail Wasserstein error relative to its QR-DQN and IQN counterparts, but its upper-tail error remains large and highly variable across seeds. The results therefore show that improved global reconstruction does not necessarily imply proportionally improved tail reconstruction.

Table 7: Heavy-Tail state-0 reconstruction diagnostics (mean $\pm$ standard deviation over 10 seeds). Tail diagnostics complement global Wasserstein error by measuring reconstruction in the rare-event regions of the return law.

Agent	Objective	W_1	Tail qMAE	Lower-tail error	Upper-tail error	Pred./true std.
QR-DQN	Quantile Huber	1.160 ± 0.161	3.365 ± 0.218	2.528 ± 0.370	4.202 ± 0.521	0.606 ± 0.043
QR-DQN	MMD	0.588 ± 0.157	2.369 ± 0.298	0.263 ± 0.220	4.475 ± 0.585	0.781 ± 0.034
QR-DQN	Sinkhorn	1.654 ± 0.101	5.647 ± 0.225	1.819 ± 0.362	9.475 ± 0.694	0.309 ± 0.014
IQN	Quantile Huber	1.069 ± 0.168	2.761 ± 0.268	2.361 ± 0.405	3.161 ± 0.645	0.654 ± 0.035
IQN	MMD	0.960 ± 0.264	2.629 ± 0.486	0.560 ± 0.388	4.698 ± 1.155	0.659 ± 0.063
IQN	Sinkhorn	1.493 ± 0.109	5.561 ± 0.408	1.736 ± 0.455	9.386 ± 0.959	0.352 ± 0.035
FQF	Quantile Huber	1.070 ± 0.326	1.596 ± 0.441	1.165 ± 0.754	2.028 ± 1.115	0.884 ± 0.065
FQF	MMD	1.233 ± 0.324	3.405 ± 0.851	2.555 ± 1.019	4.254 ± 1.937	0.926 ± 0.149
FQF	Sinkhorn	1.014 ± 0.333	3.111 ± 1.835	0.953 ± 0.678	5.268 ± 3.461	1.055 ± 0.466

B.3 Sparse-Reward Diagnostics

The Sparse regime contains both a discrete zero-return component and a continuous nonzero component. This makes the regime useful for distinguishing global reconstruction from point-mass reconstruction. QR-DQN-MMD and IQN-MMD achieve low global Wasserstein error and standard-deviation ratios near one, but they do not exactly recover the true zero-return mass. Quantile Huber regression remains more strongly underdispersed. FQF shows a different objective ordering. MMD substantially improves Sparse reconstruction relative to FQF Quantile Huber, whereas weighted Sinkhorn is weaker and more variable. These results indicate that point-mass structure is not fully characterized by global support or moment diagnostics alone.

Table 8: Sparse state-0 reconstruction diagnostics (mean ± standard deviation over 10 seeds). The Sparse return law contains a discrete zero-return mass together with a continuous component, allowing global reconstruction and point-mass reconstruction to be evaluated separately.

Agent	Objective	W_1	Pred./true std.	Pred. $P(Z=0)$	True $P(Z=0)$	Zero-mass error	Zero-plateau MAE
QR-DQN	Quantile Huber	1.129 ± 0.032	0.449 ± 0.012	0.000 ± 0.000	0.107 ± 0.000	0.107 ± 0.000	2.662 ± 0.151
QR-DQN	MMD	0.316 ± 0.098	0.949 ± 0.038	0.047 ± 0.031	0.107 ± 0.000	0.060 ± 0.031	3.680 ± 0.516
QR-DQN	Sinkhorn	2.743 ± 0.395	0.448 ± 0.065	0.053 ± 0.055	0.107 ± 0.000	0.068 ± 0.033	1.119 ± 0.559
IQN	Quantile Huber	1.110 ± 0.034	0.461 ± 0.011	0.000 ± 0.000	0.107 ± 0.000	0.107 ± 0.000	2.659 ± 0.120
IQN	MMD	0.366 ± 0.096	0.858 ± 0.042	0.057 ± 0.023	0.107 ± 0.000	0.051 ± 0.023	3.177 ± 1.341
IQN	Sinkhorn	3.033 ± 0.643	0.167 ± 0.137	0.353 ± 0.443	0.107 ± 0.000	0.288 ± 0.413	0.868 ± 0.825
FQF	Quantile Huber	0.865 ± 0.155	0.679 ± 0.067	0.000 ± 0.000	0.107 ± 0.000	0.107 ± 0.000	2.458 ± 0.280
FQF	MMD	0.541 ± 0.141	1.168 ± 0.156	0.004 ± 0.012	0.107 ± 0.000	0.103 ± 0.012	2.836 ± 1.127
FQF	Sinkhorn	1.178 ± 0.586	1.369 ± 0.494	0.022 ± 0.019	0.107 ± 0.000	0.086 ± 0.019	2.842 ± 0.916

B.4 Distribution Drift

All nine primary objective–representation combinations fail the benchmark recovery criterion after the abrupt reward-distribution shift. Final state-s_0 Wasserstein errors lie in a narrow range of approximately 19–21. The similarity across methods is notable because the same objectives behave very differently in stationary regimes. Under Drift, however, architecture and objective differences are dominated by the persistent replay mismatch created by retaining pre-drift samples throughout post-drift learning. The reported pre-drift summary metric aggregates the configured pre-drift evaluation window rather than using only the final evaluation immediately before the change. Post-drift area under the error curve is expressed in units of Wasserstein error multiplied by environment transitions.

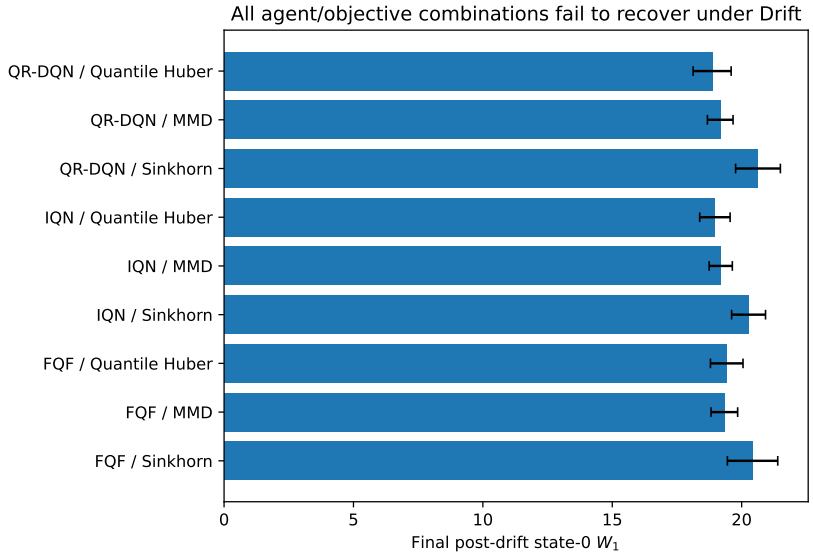

Figure 2: Final state-s_0 Wasserstein-1 error after distribution drift. All nine configurations fail the recovery criterion under the persistent stale-replay protocol.

Adaptation lag is undefined for all primary configurations because none satisfies the benchmark recovery criterion within the evaluation horizon.

Table 9: Distribution-Drift state-0 diagnostics over 10 seeds. Pre-drift error summarizes the configured pre-drift window; post-drift AUC is Wasserstein error integrated over environment transitions. Recovered is the fraction of seeds satisfying the benchmark recovery criterion.

Agent	Objective	Pre-drift	Peak post-drift	Final W_1	Recovery threshold	Recovered	Post-drift AUC
QR-DQN	Quantile Huber	5.538 ± 0.186	28.705 ± 0.389	18.859 ± 0.735	6.922 ± 0.233	0.00	760873 ± 10502
QR-DQN	MMD	6.531 ± 0.381	27.848 ± 0.413	19.172 ± 0.495	8.164 ± 0.476	0.00	718834 ± 21357
QR-DQN	Sinkhorn	5.600 ± 0.229	28.753 ± 0.311	20.631 ± 0.865	7.000 ± 0.286	0.00	780765 ± 12851
IQN	Quantile Huber	5.559 ± 0.110	28.615 ± 0.309	18.967 ± 0.587	6.949 ± 0.137	0.00	758452 ± 12390
IQN	MMD	5.768 ± 0.399	28.634 ± 0.708	19.192 ± 0.447	7.210 ± 0.499	0.00	758365 ± 19566
IQN	Sinkhorn	5.588 ± 0.189	28.626 ± 0.379	20.267 ± 0.655	6.985 ± 0.237	0.00	770047 ± 13447
FQF	Quantile Huber	5.413 ± 0.264	28.837 ± 0.328	19.423 ± 0.629	6.766 ± 0.330	0.00	775605 ± 18953
FQF	MMD	6.369 ± 0.535	28.466 ± 0.482	19.331 ± 0.514	7.961 ± 0.669	0.00	731666 ± 26437
FQF	Sinkhorn	5.916 ± 0.217	28.429 ± 0.428	20.422 ± 0.973	7.395 ± 0.272	0.00	763272 ± 16038

B.5 OOD Reconstruction

The OOD scale-shift condition provides a controlled example in which expected-value error does not reveal the principal reconstruction failure. For QR-DQN Quantile Huber, mean error is effectively unchanged between the in-distribution and scale-up conditions, while Wasserstein error increases from approximately 0.80 to 4.34. IQN quantile Huber exhibits the same phenomenon. MMD reduces the in-distribution error but still undergoes a large increase under scale shift. FQF-MMD behaves differently because its in-distribution prediction is already overdispersed; the broader OOD target therefore happens to align more closely with its learned distribution. The scale-shift results should therefore be interpreted as zero-shot distribution reconstruction rather than contextual generalization: the agent receives no explicit regime variable and performs no adaptation.

Table 10: Zero-shot OOD state-0 Wasserstein-1 reconstruction error after training on the ID Gaussian environment (mean $\pm$ standard deviation over 10 seeds; lower is better).

Agent	Objective	ID	Location +	Location -	Scale +	Combined
QR-DQN	Quantile Huber	0.801 ± 0.122	19.157 ± 0.470	19.090 ± 0.470	4.337 ± 0.043	19.157 ± 0.470
QR-DQN	MMD	0.242 ± 0.141	19.092 ± 0.276	19.155 ± 0.276	3.611 ± 0.050	19.092 ± 0.276
QR-DQN	Sinkhorn	0.595 ± 0.245	18.897 ± 0.571	19.350 ± 0.571	3.987 ± 0.086	18.897 ± 0.571
IQN	Quantile Huber	0.786 ± 0.080	19.182 ± 0.414	19.065 ± 0.414	4.335 ± 0.038	19.182 ± 0.414
IQN	MMD	0.374 ± 0.180	19.114 ± 0.416	19.133 ± 0.416	3.680 ± 0.103	19.114 ± 0.416
IQN	Sinkhorn	0.583 ± 0.398	18.952 ± 0.680	19.295 ± 0.680	3.843 ± 0.083	18.952 ± 0.680
FQF	Quantile Huber	0.553 ± 0.231	18.839 ± 0.420	19.408 ± 0.420	3.704 ± 0.156	18.839 ± 0.420
FQF	MMD	1.883 ± 0.529	19.415 ± 0.374	18.833 ± 0.374	1.806 ± 0.561	19.415 ± 0.374
FQF	Sinkhorn	1.270 ± 0.427	19.477 ± 0.566	18.770 ± 0.566	2.539 ± 0.374	19.477 ± 0.566

For the positive location intervention used here, the predicted and target return distributions are sufficiently separated that the one-dimensional W_1 coupling is monotone with a constant sign. In this regime, W_1 reduces to the absolute difference in expectations. Because the Combined intervention preserves the same shifted mean as Location+, its additional scale change does not alter the reported W_1, explaining the identical entries in those columns.

B.6 Probability-Location Controls

FQF and IQN differ from QR-DQN partly in how probability locations are selected. FQF adaptively learns fraction locations, whereas IQN samples fractions during training. To test whether these mechanisms account for the representation–objective interaction in the primary study, we conducted a targeted control over the four stationary regimes. For FQF, fraction learning was disabled and the fraction proposal was retained at uniform interval masses of $1/N$. For IQN, independently sampled training fractions were replaced by the deterministic midpoint grid

$$\tau_i = \frac{i - \frac{1}{2}}{N}, \qquad i = 1, \ldots, N.$$

Table 11: Probability-location controls on the four stationary SDRB regimes. Entries report final state-s_0 Wasserstein-1 reconstruction error (mean $\pm$ standard deviation; lower is better). Standard IQN samples fractions during training, whereas IQN Fixed uses a deterministic midpoint grid. Standard FQF adaptively learns fraction locations, whereas FQF Fixed retains uniform interval masses. QR-DQN is included as the fixed-quantile baseline from the primary matrix. All reported control conditions use the same ten benchmark seeds (123–132).

Objective	Representation	Gaussian	Mixture	Heavy Tail	Sparse
Quantile Huber	QR-DQN	0.801 ± 0.122	1.042 ± 0.069	1.160 ± 0.161	1.129 ± 0.032
	IQN	0.786 ± 0.080	1.065 ± 0.090	1.069 ± 0.168	1.110 ± 0.034
	IQN Fixed	0.821 ± 0.089	1.066 ± 0.069	1.142 ± 0.154	1.127 ± 0.044
	FQF	0.553 ± 0.231	0.903 ± 0.269	1.070 ± 0.326	0.865 ± 0.155
	FQF Fixed	0.816 ± 0.086	1.061 ± 0.069	1.141 ± 0.158	1.125 ± 0.042
MMD	QR-DQN	0.242 ± 0.141	0.431 ± 0.245	0.588 ± 0.157	0.316 ± 0.098
	IQN	0.374 ± 0.180	0.667 ± 0.278	0.960 ± 0.264	0.366 ± 0.096
	IQN Fixed	0.304 ± 0.229	0.392 ± 0.260	0.582 ± 0.212	0.563 ± 0.485
	FQF	1.883 ± 0.529	2.128 ± 1.009	1.233 ± 0.324	0.541 ± 0.141
	FQF Fixed	0.294 ± 0.196	0.380 ± 0.260	0.568 ± 0.212	0.310 ± 0.185
Sinkhorn	QR-DQN	0.595 ± 0.245	0.798 ± 0.241	1.654 ± 0.101	2.743 ± 0.395
	IQN	0.583 ± 0.398	0.733 ± 0.513	1.493 ± 0.109	3.033 ± 0.643
	IQN Fixed	0.602 ± 0.230	0.858 ± 0.232	1.700 ± 0.132	2.351 ± 0.539
	FQF	1.270 ± 0.427	1.030 ± 0.531	1.014 ± 0.333	1.178 ± 0.586
	FQF Fixed	0.602 ± 0.369	0.747 ± 0.264	1.711 ± 0.188	2.158 ± 0.335

All other representation, objective, replay, optimization, target-update, and training settings were retained. Each control condition uses the same ten benchmark seeds (123–132) as the primary factorial study.

The intervention has strongly objective-dependent effects. Under Quantile Huber, fixing FQF fractions increases reconstruction error in all four stationary regimes, whereas under MMD the same intervention decreases error in all four. Sinkhorn produces a heterogeneous result: fixed FQF improves Gaussian and Mixture reconstruction but degrades Heavy-Tail and Sparse reconstruction. Fixed-grid IQN is similarly nonuniform. Its Quantile Huber performance changes comparatively little, while under MMD the fixed grid improves Gaussian, Mixture, and Heavy Tail but degrades Sparse; under Sinkhorn, it degrades Gaussian, Mixture, and Heavy Tail while improving Sparse.

The controls therefore identify probability-location selection as a contributor to reconstruction performance without reducing the primary interaction to that mechanism alone. In several conditions, fixed-grid IQN and fixed-fraction FQF attain similar reconstruction errors, but this similarity is not universal across return laws. More generally, the effect of fixing probability locations depends jointly on the representation, distributional objective, and target return-law geometry.

B.7 Representation Diagnostics

Quantile-crossing rate is informative about organization of the learned outputs but should not be interpreted as a generic measure of distribution quality. MMD is permutation invariant and can achieve low Wasserstein reconstruction error despite high crossing rates. For example, successful MMD runs frequently exhibit crossing rates near one half because the objective does not impose an ordering on the empirical atoms. FQF fraction diagnostics include the entropy of the learned interval probabilities, minimum and maximum interval probability, and mean absolute deviation from the uniform allocation. These diagnostics are especially useful for identifying objective-dependent interactions between the learned quantile surface and the FQF fraction-update rule.

C Objective and Agent Calibration

All objective- and agent-specific hyperparameters used in the primary benchmark matrix were selected using small diagnostic experiments before the full factorial evaluation. Once selected, these

settings were frozen across benchmark regimes and random seeds. We did not tune hyperparameters separately for individual worlds.

C.1 Quantile Huber Regression

Quantile Huber regression serves as the conventional baseline objective. Its configuration was frozen before introducing alternative objectives and was retained unchanged throughout the benchmark matrix. The resulting QR-DQN baseline exhibits a characteristic stationary underdispersion pattern that is reproduced under IQN. Because this behavior is stable across ten seeds and multiple stationary regimes, we treat it as an empirical property of the frozen baseline rather than a tuning failure.

C.2 MMD Kernel Selection

We initially evaluated a single-scale RBF MMD objective with bandwidth given by the detached median positive target pairwise distance. Although this objective was numerically stable, its Gaussian reconstruction was substantially worse and more variable than desired.

We therefore evaluated a multi-scale variant using the common base bandwidth with multiplicative scales

$$0.25, 0.5, 1, 2, 4. \tag{9}$$

This configuration substantially improved Gaussian reconstruction and was frozen for the primary matrix. The bandwidth heuristic remains based only on the detached target atom geometry and is not optimized by gradient descent.

C.3 Sinkhorn Entropy Calibration

The Sinkhorn objective uses the debiased entropic divergence

$$S_\varepsilon(P, Q) = \mathrm{OT} * \varepsilon(P, Q) - \frac{1}{2}\mathrm{OT} * \varepsilon(P, P) - \frac{1}{2}\mathrm{OT}_\varepsilon(Q, Q). \tag{10}$$

We performed a small diagnostic sweep over the scaling used to convert the detached target median pairwise ground cost into the entropic regularization coefficient. The final configuration uses

$$\varepsilon = \max\left(10^{-5}, 0.003 \operatorname*{median}_{i \neq j} |y_i - y_j|\right).$$

We use 50 Sinkhorn iterations for every primary experiment. This setting was retained throughout the matrix even though Sinkhorn is substantially more computationally expensive than Quantile Huber regression and MMD, because changing the iteration count after partial evaluation would have confounded comparisons across agents.

C.4 FQF Fraction Learning Rate

The nominal FQF fraction-network learning rate of 2.5×10^{-9} left the learned fraction probabilities effectively uniform under the SDRB Gaussian diagnostic condition.

For $N = 51$ intervals, uniform probability is approximately

$$\frac{1}{51} \approx 0.01961, \tag{12}$$

and the maximum entropy is

$$\log 51 \approx 3.9318. \tag{13}$$

At 2.5×10^{-9}, the final entropy remained approximately 3.9318, and the mean absolute deviation from the uniform allocation was only approximately 3.5×10^{-5}. We therefore performed a logarithmic diagnostic sweep. Increasing the learning rate produced progressively stronger fraction adaptation, but larger values also produced increasingly concentrated interval allocations. A refined diagnostic included 10^{-5}, 5×10^{-5}, and 10^{-4}. At 10^{-5}, the Gaussian diagnostic produced

meaningful but moderate adaptation: entropy decreased to approximately 3.914, the minimum interval probability was approximately 0.0069, the maximum approximately 0.0334, and mean absolute deviation from uniform approximately 0.00275. Larger rates produced substantially stronger concentration.

We therefore selected

$$\eta_{\text{fraction}} = 10^{-5} \tag{14}$$

for the primary FQF experiments.

The selection criterion was mechanism activation and stability rather than optimization of the final single-seed Wasserstein score. Specifically, we selected the lowest tested learning rate that produced substantial fraction adaptation without the highly concentrated allocations observed at larger rates.

C.5 Weighted Objectives for FQF

QR-DQN and IQN produce equal-weight empirical distributions. FQF instead associates each quantile value with interval probability

$$w_i = \tau_{i+1} - \tau_i. \tag{15}$$

Applying equal-weight MMD or Sinkhorn directly to FQF atoms would therefore compare the wrong empirical measure. We extended both objectives to support nonuniform probability weights.

For weighted MMD,

$$\mathrm{MMD}^2 = \mathbf{a}^\top K_{XX}\mathbf{a} + \mathbf{b}^\top K_{YY}\mathbf{b} - 2\mathbf{a}^\top K_{XY}\mathbf{b}, \tag{16}$$

where $\mathbf{a}$ and $\mathbf{b}$ are normalized probability weights.

For Sinkhorn, the same weights define the source and target transport marginals in the P–Q, P–P, and Q–Q terms. The value-objective update receives detached interval probabilities. The fraction-proposal network is updated only by the canonical FQF fraction objective. This separation is verified explicitly in Appendix D.

This decomposition preserves the canonical separation between fraction learning and value learning, but it also creates a potential compatibility issue when the value objective is permutation invariant. The canonical FQF fraction surrogate is defined over indexed quantile locations and implicitly relies on a locally ordered quantile surface, whereas MMD and Sinkhorn compare empirical measures without directly penalizing permutation of return particles. Consequently, gradient isolation alone does not guarantee that the two optimization components are geometrically compatible. The crossing and fraction-concentration diagnostics reported in Appendix D are therefore interpreted as evidence of a possible objective–fraction-update interaction, not as evidence of a software gradient leak.

D Implementation Validation

Because the benchmark compares multiple objectives and distributional representations through a shared training pipeline, we performed targeted regression and gradient-isolation tests before running the full matrix.

D.1 Objective Injection

Distributional objectives are injected through a common objective interface rather than implemented as separate learners. The original Quantile Huber path was preserved and tested for exact regression equivalence after introducing objective injection. This design reduces the likelihood that objective comparisons inadvertently change unrelated replay, target-network, optimization, or evaluation behavior.

D.2 Weighted Empirical Measures

For MMD and Sinkhorn, the original equal-weight computation path is retained when no probability weights are supplied. Weighted support is activated only for representations such as FQF that

induce unequal empirical masses. Regression tests verify that the equal-weight path is unchanged relative to the frozen QR-DQN and IQN implementations. Weighted-objective tests additionally verify normalization and validity of the supplied probability masses.

D.3 FQF Gradient Isolation

FQF contains two conceptually separate optimization processes: learning return values and learning fraction locations. For weighted empirical-measure objectives, this creates a potential failure mode in which probability weights could unintentionally allow the value objective to update the fraction-proposal network. We explicitly test both gradient pathways. First, a weighted MMD or Sinkhorn value-objective update produces nonzero gradients for the value network while producing no gradient in the fraction-proposal network. Second, the canonical FQF fraction surrogate produces nonzero gradients for the fraction-proposal network while producing no gradient in the value network. Both conditions are verified for weighted MMD and weighted Sinkhorn. These tests ensure that objective comparisons do not alter the intended FQF optimization decomposition.

D.4 Evaluation Compatibility

All representations are evaluated through a common 51-point interface. QR-DQN directly returns its 51 learned atoms. IQN is evaluated using a deterministic grid of 51 midpoint quantile fractions rather than sampled evaluation fractions. FQF retains its learned adaptive representation for action-value computation and training while exposing a common evaluation interface for generic reconstruction metrics. This evaluation design allows Wasserstein, CDF, support, moment, and specialized benchmark metrics to be computed consistently across representations.

D.5 FQF Fraction Diagnostics

We monitor fraction entropy, minimum interval probability, maximum interval probability, and mean absolute deviation from the uniform allocation throughout FQF training. Under quantile Huber regression, fraction adaptation remains moderate. Under weighted MMD, fraction allocations can become substantially more concentrated, particularly on Gaussian and Sparse regimes. Weighted Sinkhorn also induces concentration, although typically less severely. We do not treat concentration alone as an implementation failure. Smoke-test trajectories show that these allocations emerge gradually during learning rather than appearing at initialization, and gradient-isolation tests confirm that the weighted value objective does not directly update the fraction network. We therefore interpret these patterns as objective–representation interactions mediated by the learned quantile surface.

D.6 Oracle Reproducibility

High-precision Monte Carlo ground-truth distributions use a fixed oracle seed and are cached independently of training seeds and objectives. This ensures that every learner is evaluated against the same numerical target. Persistent oracle caching also avoids repeatedly recomputing expensive Heavy-Tail and Sparse return targets during the large factorial benchmark.

D.7 Benchmark Auditing

Before aggregation, benchmark outputs are checked for schema consistency, valid benchmark and split names, finite metric values, duplicate records, registry consistency, primary-metric flags, and case/regime conventions. The final paper aggregation additionally verifies that every selected agent–objective pair contains all ten expected seeds and that the complete 3×3 matrix is present before paper tables are produced.

E Diagnostic Study of Density-Based Divergences

We additionally investigated three density-based discrepancy measures as candidate objectives: Jensen–Shannon divergence, squared Hellinger distance, and Rényi divergence. These objectives were not included in the primary factorial benchmark because the diagnostic study below revealed a systematic failure associated with the density approximation on the Sparse benchmark regime.

E.1 MOTIVATION

The primary objectives in the main paper operate directly on empirical return measures. Quantile Huber regression operates through quantile targets, MMD compares empirical measures through kernels, and Sinkhorn compares them through optimal transport. In contrast, Jensen–Shannon, Hellinger, and Rényi divergences are naturally expressed in terms of probability densities. The learned return representations in our experiments are empirical measures of the form

$$\widehat{P} = \sum_i w_i \delta_{x_i}. \tag{17}$$

Because independently learned atoms almost never occupy identical support locations, directly applying a discrete density divergence would be degenerate. We therefore constructed differentiable Gaussian kernel density estimates from the empirical atoms before evaluating the divergence.

E.2 KDE APPROXIMATION

Given atoms and probability weights, we construct a Gaussian KDE and evaluate the resulting density on a one-dimensional quadrature grid. The bandwidth is determined from the detached target median positive pairwise atom distance. The diagnostic configuration uses a bandwidth scale of 1, minimum bandwidth 10^{-3}, 129 quadrature points, and a grid extending four bandwidths beyond the combined predicted and target atom range. The resulting densities are numerically normalized on the finite quadrature grid before evaluating the divergence. Grid endpoints and bandwidth selection are detached from gradient propagation; gradients flow through the KDE density values and therefore through the predicted atom locations.

For Jensen–Shannon divergence,

$$\mathrm{JS}(P, Q) = \frac{1}{2}\,\mathrm{KL}(P|M) + \frac{1}{2}\,\mathrm{KL}(Q|M), \qquad M = \frac{1}{2}(P + Q). \tag{18}$$

Squared Hellinger distance is

$$H^2(P, Q) = \frac{1}{2}\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx. \tag{19}$$

Rényi divergence is

$$D_\alpha(P|Q) = \frac{1}{\alpha - 1}\log \int p(x)^\alpha q(x)^{1-\alpha}, dx. \tag{20}$$

E.3 GAUSSIAN DIAGNOSTIC

Single-seed Gaussian diagnostics initially suggested that all three families were plausible. Jensen–Shannon and Hellinger produced finite gradients and competitive reconstruction across QR-DQN, IQN, and FQF. Their optimization behavior was substantially better conditioned than the initial Rényi configuration. Rényi divergence with $\alpha = 2$ exhibited large gradient spikes and unstable reconstruction trajectories. We therefore performed a small order-parameter diagnostic using $\alpha \in 0.75, 1.25, 1.5, 2$. Values above one remained poorly conditioned. The $\alpha = 0.75$ configuration was substantially more stable and was selected for additional diagnostic testing. We did not perform objective-specific optimizer tuning.

E.4 HEAVY-TAIL DIAGNOSTIC

Jensen–Shannon and Hellinger remained trainable on the Heavy-Tail regime. In particular, FQF produced substantially better Heavy-Tail reconstruction under both objectives than its QR-DQN and IQN counterparts. Rényi-0.75 also trained, although it remained more gradient-sensitive and produced weaker tail reconstruction. If evaluation had stopped at Gaussian and Heavy Tail, Jensen–Shannon and Hellinger would have appeared to be viable additions to the primary objective set.

E.5 Sparse-Reward Failure

The Sparse regime produced a qualitatively different result. Under Jensen–Shannon, IQN and FQF converged to predictions with effectively all represented probability mass at zero. Their predicted zero mass approached one despite the true zero-return mass being approximately 0.107. Their predicted standard deviations approached zero and their final state-s_0 Wasserstein errors remained near 3.82. QR-DQN did not collapse completely but still substantially overestimated the zero-return mass and remained near the same high Wasserstein error. Squared Hellinger reproduced essentially the same failure across all three representations. Rényi-0.75 also failed to reconstruct the Sparse return distribution and exhibited additional gradient-conditioning problems. These failures are especially informative because the Sparse return law is mixed discrete–continuous. The Gaussian KDE bridge necessarily converts this structure into a purely continuous density before applying the divergence. The resulting density geometry can therefore strongly reward concentration around the dominant zero component while failing to preserve the nonzero return structure.

E.6 Exclusion from the Primary Matrix

We therefore excluded KDE-based density divergences from the primary factorial experiment. This decision was not based on their failure to outperform the existing objectives. Poor performance is itself a valid benchmark result. Instead, the diagnostic experiments revealed that their behavior depended critically on an auxiliary density-estimation layer whose interaction with mixed discrete–continuous targets created a systematic structural pathology. Including these objectives in the main matrix would therefore conflate the geometry of the divergence with an additional KDE modeling choice. The diagnostic study also provides a small demonstration of why benchmark breadth matters. Gaussian evaluation alone would have suggested that Jensen–Shannon and Hellinger were viable. Heavy-Tail evaluation would not have changed that conclusion. The Sparse regime exposed a failure that was not visible in either of those environments. Future work could investigate divergences designed directly for mixed measures or hybrid objectives that explicitly separate discrete and continuous components. Accordingly, these results should be interpreted as a failure of the tested *KDE-mediated implementations* on the mixed discrete–continuous Sparse target, not as evidence that Jensen–Shannon, Hellinger, or Rényi divergences are intrinsically unsuitable for distributional reinforcement learning.