Optimal Transport Meets Reinforcement Learning: A Survey

Anonymous authors
Paper under double-blind review

Abstract

Reinforcement learning (RL) algorithms frequently compare probability distributions: the states a policy visits versus those an expert visits, the actions a learned policy takes versus those in a dataset, or the transitions a model predicts versus those the environment produces. Standard divergences such as Kullback–Leibler and Jensen–Shannon can fail when these distributions overlap weakly—a common situation in imitation learning, offline RL, and deployment under distribution shift. Optimal transport (OT) offers an alternative by measuring the cost of *moving* probability mass from one distribution to another under a ground cost that encodes task geometry. This often provides informative gradients even under support mismatch. This survey covers how OT is used inside RL objectives and algorithms. For each method, we identify: the role OT plays (reward signal, regulariser, constraint, ambiguity set, or model loss), the distributions compared (occupancies, action conditionals, trajectory embeddings, transition kernels, or return distributions), the OT formulation used W_1, W_2, (W_1, W_2, entropy-regularised/Sinkhorn, Gromov–Wasserstein, or partial/unbalanced variants), and the treatment of temporal structure. We summarise practical guidance on cost design, regularisation tuning, and numerical stability, and highlight open problems including scalable trajectory-level transport, principled handling of mass mismatch, and theoretical guarantees for OT-regularised RL.

1 Introduction

Reinforcement learning (RL) studies how an agent should act in an uncertain environment to maximise long-term reward. The agent interacts with the environment, observes states, takes actions, receives rewards, and updates its behaviour (Sutton & Barto, 1998). This framework covers problems from robotics and control (Kober et al., 2013; Schulman et al., 2017; Lillicrap et al., 2015) to resource allocation (Ye et al., 2019; Luong et al., 2019; Cui et al., 2019) and dialogue systems (Li et al., 2016; Su et al., 2016; Papangelis et al., 2019), but it also exposes recurring difficulties: limited data (Levine et al., 2020; Sun & Wu, 2023; Zhu et al., 2025), compounding error over time (Talvitie, 2017; Janner et al., 2019; Jiang, 2024), and the need to generalise under distribution shift (Cobbe et al., 2019; Wang et al., 2020; Fujimoto et al., 2024).

Although RL is often presented as expected-return maximisation, many RL algorithms are really comparing and shaping *distributions* (Zhang et al., 2020; Papagiannis & Li, 2022; Dadashi et al., 2021a). For example, a policy induces a distribution over visited states and actions, namely its *occupancy measure*. Informally, this is the policy's statistical "footprint": where it tends to spend time, and which actions it tends to take there. Several familiar RL problems can be phrased in terms of distribution comparisons. In imitation learning, the question is how far the learner's footprint is from the expert's (Xiao et al., 2019; Luo et al., 2023; Chang et al., 2023). In policy optimisation, the learning stability is influenced by how far the learned policy drifts from the policy from the previous iteration (Song et al., 2023; Pacchiano et al., 2020; Pfau et al., 2025). In offline RL, the issue is how far the learned policy deviates from the behaviour policy that generated the data (Wu et al., 2019; Omura et al., 2025; Asadulaev et al., 2024). In robust control, one asks how much the transition law may move away from a nominal model (Abdullah et al., 2019; Hou et al., 2020; Queeney et al., 2024). In curriculum learning, one needs a principled way to move from one task distribution to another (Klink et al., 2022; Huang et al., 2022). Across all of these settings, progress depends on having a discrepancy that says not only whether two distributions differ, but how they differ in ways that matter for control.

One choice is to measure discrepancy using f-divergences such as Kullback–Leibler (KL) or Jensen–Shannon (JS). These compare densities pointwise: at each state or action, they ask how the two probability densities differ. This works well when distributions overlap substantially, but becomes problematic when they do not

(Birrell et al., 2022). In imitation learning, for instance, early in training the learner often visits states the expert never visited. A JS discriminator may then saturate and provide almost no gradient, and a KL term may be undefined or dominated by regions where one density is near zero (Arjovsky & Bottou, 2017). The learning signal degrades precisely when guidance is most needed.

Optimal transport (OT) takes a different approach. Rather than comparing densities pointwise, OT asks: what is the minimum effort required to *move* probability mass from one distribution to another? A useful mental picture is to imagine reshaping one pile of sand into another. The amount of work depends both on how much sand must be moved and on how far it must move. In OT, that movement is quantified by a *ground cost*, which specifies how expensive it is to transport mass between any two points (Santambrogio, 2015). The key consequence is that OT remains informative even when distributions have disjoint support. If a learner state is far from all expert states, OT quantifies *how far* in the geometry defined by the ground cost rather than returning an uninformative or saturated value. This geometric perspective also exposes a modelling lever absent from f-divergences: the ground cost encodes domain knowledge about what it means for two states, actions, or trajectories to be close. A robotic manipulation task might use kinematic distance; a navigation task might use geodesic distance; and a high-dimensional observation space might use distance in a learned feature embedding. The choice of cost directly shapes what the algorithm treats as similar or different.

These properties have led to OT being adopted across many areas of RL. In imitation learning, Wasserstein distances match learner and expert occupancies, either adversarially through dual critics (Xiao et al., 2019) or through transport plans that assign rewards offline (Luo et al., 2023). In policy optimisation, Wasserstein constraints replace KL-based trust regions so that small updates respect action geometry rather than only likelihood ratios (Song et al., 2023). In robustness, Wasserstein balls around nominal dynamics define ambiguity sets for worst-case optimisation (Yang, 2017). In world modelling, Wasserstein losses on one-step transitions can be related to worst-case value error over Lipschitz test functions (Asadi et al., 2018). In transfer and curricula, OT, and Gromov–Wasserstein when spaces differ, aligns distributions across domains and interpolates between task distributions in a geometry-aware way (Liu et al., 2024; Klink et al., 2022). In offline RL, transport-based constraints keep learned policies close to the behaviour policy that generated the data (Wu et al., 2019). The technical details differ from section to section, but the recurring theme is simple: OT gives RL algorithms a way to compare distributions using geometry without requiring overlapping support.

Despite this breadth, the literature is fragmented. Methods developed for imitation may not cite related ideas in robustness, and techniques for trust regions may not connect to distributional critics. Without a unifying view, it is easy to treat OT as a black-box distance and overlook the modelling choices that determine success or failure: a mis-specified ground cost, poorly scaled entropic regularisation, an unstable dual critic, or a coupling that ignores temporal structure. This survey is selective rather than exhaustive; it includes works where OT appears explicitly in the objective, constraint, solver, ambiguity set, or evaluation metric.

This survey therefore focuses on OT *as it appears inside RL objectives and algorithms*. Throughout, we keep returning to four concrete questions: what OT is doing inside the objective, which distributions are being compared, which OT formulation is used, and how temporal structure is handled. Those questions organise the survey (Figure 1) and make it easier to compare methods that are usually discussed in separate literatures. The recurring roles are reward signal, regulariser, trust-region constraint, ambiguity set, model loss, and evaluation metric.

After preliminaries on RL and OT (Section 2), we cover imitation learning and reward inference (Section 3), policy optimisation and regularisation (Section 4), model-based RL (Section 5), offline RL (Section 6), robustness, safety, and transfer (Section 7), and other applications including multi-policy and multi-agent settings (Section 8). We conclude with practical guidance and open problems (Section 9).

2 Preliminaries

This section fixes the notation and core definitions used throughout the survey. Section 2.1 introduces the MDP framework, occupancy measures, and the distributional objects that OT will act on. Section 2.2 then

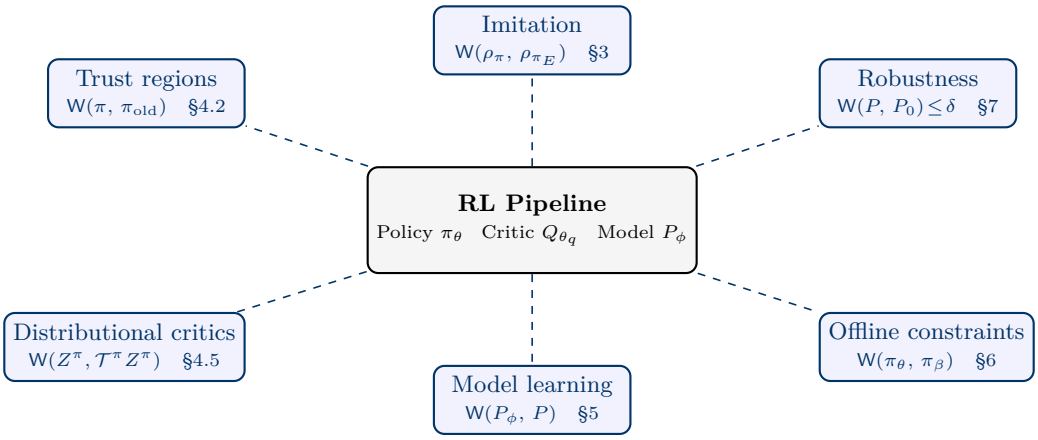

Figure 1: Where optimal transport enters the RL pipeline. Each spoke represents a distinct role that OT plays, with the representative objective and section reference. Additional roles not shown include safety constraints (Sections 7.2, 7.3), preference alignment (Section 7.4), transfer and curricula (Sections 7.6, 8.3), and multi-policy/multi-agent diversity (Section 8).

reviews the OT concepts, including Wasserstein distances, entropic regularisation, Gromov–Wasserstein, and partial/unbalanced variants, that occur in later sections.

2.1 Reinforcement learning in brief

Optimal transport enters RL through *probability measures induced by a policy*: per-state action distributions $\pi(\cdot \mid s)$, discounted occupancies ρ_π, and trajectory laws P_π or their pushforwards under an embedding map Φ. This subsection fixes notation for these objects and gives the minimum intuition needed for the later sections. The key idea is simple: OT does not act directly on policy parameters, but on the distributions the policy generates, that is, *where the policy places mass* and *how that mass moves through the dynamics*.

Sutton & Barto (1998) provides the foundation for RL. The environment is usually modelled as a discounted Markov decision process (MDP)

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma, p_0),$$

where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P(\cdot \mid s, a)$ is a transition kernel on $\mathcal{S}$, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a reward function, $\gamma \in (0, 1)$ is a discount factor, and p_0 is an initial-state distribution. A (possibly stochastic) policy is a conditional distribution $\pi(\cdot \mid s)$ on $\mathcal{A}$. A trajectory can be written as

$$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \dots), \qquad r_t := R(s_t, a_t),$$

and is sampled by

$$s_0 \sim p_0, \quad a_t \sim \pi(\cdot \mid s_t), \quad s_{t+1} \sim P(\cdot \mid s_t, a_t).$$

Throughout, $R(s, a)$ denotes the expected immediate reward, r_t denotes the realised reward, and $Z^\pi(s, a)$ denotes the random discounted return induced by rewards, transitions, and policy stochasticity. It is often convenient to name the induced distribution on trajectories: let P_π denote the trajectory law generated by (p_0, π, P). We write $\mathbb{E}_\pi[\cdot]$ for expectations with respect to this induced joint distribution over trajectories (equivalently, $\mathbb{E}_{\tau \sim P_\pi}[\cdot]$), and specify the underlying sampling measure explicitly otherwise. The discounted return from time t is

$$G_t = \sum_{i=t}^{\infty} \gamma^{i-t} r_i = \sum_{i=t}^{\infty} \gamma^{i-t} R(s_i, a_i),$$

so future rewards have less weight, by a factor γ^{i-t}, compared to the current reward.

The *state-value* and *action-value* functions under a fixed policy π are

$$v^\pi(s) = \mathbb{E}_\pi[G_0 \mid s_0 = s], \qquad q^\pi(s, a) = \mathbb{E}_\pi[G_0 \mid s_0 = s, a_0 = a].$$

The *advantage* function measures the benefit of taking a in s relative to the average behaviour of π:

$$A^\pi(s,a) = q^\pi(s,a) - v^\pi(s).$$

Intuitively, $A^\pi(s,a) > 0$ signals that a is better than what π typically does in s, and $A^\pi(s,a) < 0$ the opposite. The Bellman relations simply encode one-step consistency: today's value is immediate reward plus expected discounted future value,

$$v^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\Big[R(s,a) + \gamma\,\mathbb{E}_{s'\sim P(\cdot|s,a)}[v^\pi(s')]\Big], \tag{1}$$

$$q^\pi(s,a) = R(s,a) + \gamma\,\mathbb{E}_{s'\sim P(\cdot|s,a),\,a'\sim\pi(\cdot|s')}[q^\pi(s',a')]. \tag{2}$$

Occupancy measures: a distributional "footprint" of a policy. A central object for what follows is the *discounted occupancy measure* (state–action visitation distribution)

$$\rho_\pi(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^{\,t}\,\Pr(s_t = s,\ a_t = a \mid \pi). \tag{3}$$

The prefactor $(1-\gamma)$ normalises ρ_π to a probability measure under standard measurability assumptions, so that in discrete spaces $\sum_{s,a}\rho_\pi(s,a) = 1$, and in general $\int \mathrm{d}\rho_\pi(s,a) = 1$. Informally, ρ_π tells us how often the policy visits each state–action pair after averaging over time and discounting the distant future. A useful way to interpret ρ_π (especially in continuous spaces) is via test functions: for any measurable h,

$$\int h(s,a)\,\mathrm{d}\rho_\pi(s,a) = (1-\gamma)\,\mathbb{E}_\pi\Big[\sum_{t=0}^{\infty}\gamma^t h(s_t,a_t)\Big].$$

Thus ρ_π is the discounted time-average distribution of visited state–action pairs. Its state marginal is $\rho_\pi(s) = \sum_a \rho_\pi(s,a)$ and the conditional distribution recovers the policy: $\rho_\pi(a \mid s) = \pi(a \mid s)$ whenever $\rho_\pi(s) > 0$. This is one reason occupancies are so useful: they package both *where* the agent goes and *what* it does there into a single object. The fixed-point identity

$$\rho_\pi(s,a) = (1-\gamma)\,p_0(s)\,\pi(a \mid s)\ +\ \gamma\,\mathbb{E}_{\tilde s,\tilde a\sim\rho_\pi}\big[P(s \mid \tilde s,\tilde a)\,\pi(a \mid s)\big]$$

expresses that discounted visits arise either from the start distribution (first term) or by flowing through the dynamics under π (second term). It is useful to read this as a conservation law for probability mass: some mass enters at the initial distribution, and the rest is transported forward by the dynamics and the policy. This flow perspective is the starting point for several OT-based methods, particularly those that operate on occupancy measures (Sections 3).

The flow constraint In finite MDPs, the occupancy measure is not an arbitrary distribution on $\mathcal{S}\times\mathcal{A}$: it must satisfy a conservation-of-mass constraint induced by the dynamics. Writing $\rho_\pi(s) = \sum_a \rho_\pi(s,a)$, one has the equivalent state-level balance equation

$$\rho_\pi(s) = (1-\gamma)p_0(s) + \gamma\sum_{\tilde s,\tilde a}\rho_\pi(\tilde s,\tilde a)\,P(s \mid \tilde s,\tilde a), \tag{4}$$

and then $\rho_\pi(s,a) = \rho_\pi(s)\pi(a \mid s)$. This makes clear why many OT constructions in RL are naturally stated on occupancies: changing a policy changes its occupancy by *pushing probability mass through the Markov dynamics*. In later sections, OT distances will be used either to compare two occupancies (e.g., learner vs. expert) or to constrain how far an update can move in occupancy space. For a reader seeing this for the first time, the important point is that OT is being used on a bona fide distribution that already encodes sequential behaviour.

The control objective can be written compactly in terms of ρ_π:

$$J(\pi) := \mathbb{E}_\pi[G_0] = \frac{1}{1-\gamma}\,\mathbb{E}_{(s,a)\sim\rho_\pi}[R(s,a)]. \tag{5}$$

Thus many learning objectives reduce to expectations with respect to either ρ_π (for occupancy-level matching and regularisation) or $\pi(\cdot \mid s)$ averaged over states drawn from ρ_π (for policy-space constraints).

Finite-horizon empirical measures and "bag-of-states" viewpoints. For finite-horizon problems with horizon T, an undiscounted analogue is often used. Given a single rollout $\tau = (s_0, a_0, r_0, \ldots, s_{T-1}, r_{T-1})$, the empirical *state marginal* (a "bag-of-states" measure) is

$$\hat{p} = \frac{1}{T} \sum_{t=0}^{T-1} \delta_{s_t},$$

where δ_x denotes a Dirac mass at x. This empirical measure will serve as the basic object in trajectory-level transports used for offline reward labelling. The phrase "bag of states" is meant literally: we keep track of which states appeared, but not in which order. That simplification can be helpful computationally, but it also throws away temporal information. Later sections discuss when this is acceptable and how to reintroduce temporal structure (e.g., banded/masked couplings and windowed/intent-aware costs). OT on empirical trajectory marginals is not sequence-aware by itself; temporal structure must enter through the transported object, the cost, or the admissible couplings. In continuous spaces, sums become integrals and Dirac masses become point masses in the appropriate topology; we reuse the same notation with the measure-theoretic interpretation understood.

Distributional objects used later. Beyond ρ_π and $\pi(\cdot \mid s)$, later sections also use: (i) the trajectory law P_π and its pushforwards $\mu_\pi = \Phi_\# P_\pi$ under behavioural embeddings Φ (Section 4.3); (ii) expert and behaviour policies π_E and π_β (imitation learning and offline RL); and (iii) return distributions $Z^\pi(s, a)$ in distributional RL (Section 4.5). The guiding principle is consistent: OT compares *policy-induced measures* at the level appropriate to the application.

To summarise the notation that will recur:

- $\pi(\cdot \mid s)$: policy; $P(\cdot \mid s, a)$: dynamics; p_0: initial-state distribution. They generate P_π: trajectory law.
- v^π, q^π, A^π: value, action-value, and advantage under π.
- ρ_π: discounted occupancy measure; $\rho_\pi(s) = \sum_a \rho_\pi(s, a)$; $\rho_\pi(a \mid s) = \pi(a \mid s)$.
- $\hat{p}$: empirical state marginal of a single trajectory (finite-horizon).

A more comprehensive list of notation is provided in Table 2. These objects let us state, with precision, which distributions are compared by each OT construction (e.g., occupancies ρ, empirical marginals $\hat{p}$, or action-conditionals $\pi(\cdot \mid s)$) and where the transport signal enters the RL objective.

2.2 Optimal transport in brief

We summarise the OT concepts used throughout the survey and fix notation. For background, we refer to Peyré & Cuturi (2019) for computational OT and to Villani (2008); Santambrogio (2015) for an accessible applied introduction and broader theoretical context.

Why OT. Many RL problems involve comparing distributions whose supports overlap weakly: learner versus expert occupancies early in imitation learning, learned-policy actions versus logged actions in offline RL, or transition laws under shift. Pointwise divergences such as KL can be brittle under such mismatch, whereas OT remains informative by measuring the *effort to move mass* under a ground cost that encodes task geometry. A minimal toy example is if $\mu = \delta_x$ and $\nu = \delta_y$, then $\mathsf{W}_1(\mu, \nu) = d(x, y)$ measures the distance between those two points, while $\mathrm{KL}(\mu \| \nu)$ is infinite for $x \neq y$. Generally, OT lifts the pointwise comparisons to the distribution level. As a result, OT can still provide informative gradients even when two distributions have little or no overlap in support, whereas likelihood-ratio divergences such as KL may become undefined or infinite in such settings. That is the practical reason OT often continues to provide a useful learning signal under distribution shift and consider the geometry of underlying space. However, the effect of OT highly reply on the choice of ground cost. Figure 2 illustrates the key modelling choice in the later RL sections: even with the same source and target marginals, changing the ground cost changes the optimal transport plan, and therefore changes what is considered to be a good match.

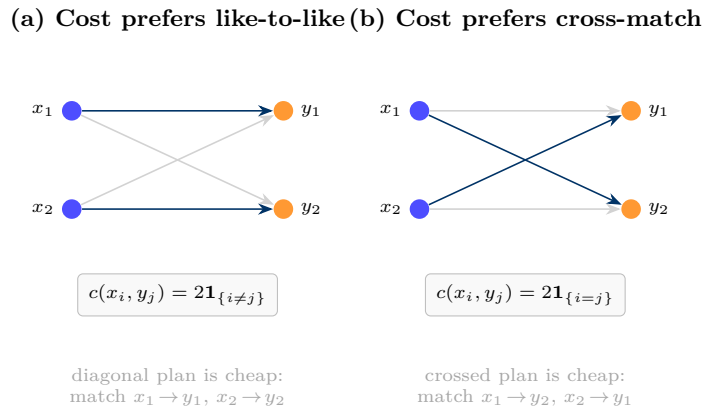

Figure 2: The ground cost determines what OT counts as a good match. In both panels the source and target marginals are the same, but the ground cost changes. Because OT minimises transport cost under that matrix, the optimal plan changes as well. This is why the ground cost is a modelling choice, not a numerical detail.

Maps, couplings, and pushforwards. Let $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ be probability measures on measurable spaces $\mathcal{X}, \mathcal{Y}$, and let $F : \mathcal{X} \to \mathcal{Y}$ be a measurable map. The *pushforward* (image) of μ through F is the measure $F_{\#}\mu \in \mathcal{P}(\mathcal{Y})$ defined by

$$(F_{\#}\mu)(B) = \mu(F^{-1}(B)), \qquad \forall\, B \subseteq \mathcal{Y} \text{ measurable.}$$

A *transport map* is a measurable F such that $F_{\#}\mu = \nu$. For instance, if μ is a distribution over states and Φ a feature map, then $\Phi_{\#}\mu$ is the induced distribution in feature space. A pushforward simply draws samples $x \sim \mu$, transforms it with F, and obtains the resulting distribution of outputs. The relaxation of a transport map is a *coupling* (transport plan) between μ and ν, which is a probability measure $\Gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ whose marginals are μ and ν. Whereas a transport map sends mass at each point to a single destination, a transport coupling can split mass across multiple destinations. This matters because in many practical problems the source mass cannot be matched by a deterministic one-to-one assignment. Writing $P_{\mathcal{X}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ and $P_{\mathcal{Y}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ for coordinate projections, the set of couplings is

$$\Pi(\mu, \nu) = \big\{\Gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : \ (P_{\mathcal{X}})_{\#}\Gamma = \mu, \ (P_{\mathcal{Y}})_{\#}\Gamma = \nu\big\}.$$

For a simple example, if $\mu = \delta_x$ and $\nu = \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}$, then no transport map exists. A coupling handles this immediately by sending mass $1/2$ from x to each y_i.

Monge and Kantorovich formulations. Given a ground cost $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, the classical Monge problem (Monge, 1781) seeks a transport map F that pushes μ to ν and minimises transport cost (Santambrogio, 2015; Peyré & Cuturi, 2019):

$$\inf_{F:\, F_{\#}\mu = \nu} \int_{\mathcal{X}} c\big(x, F(x)\big)\, \mathrm{d}\mu(x).$$

Transport maps may fail to exist (for example when mass must be split), motivating the Kantorovich relaxation (Kantorovitch, 1958):

$$\mathrm{OT}_c(\mu, \nu) := \inf_{\Gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y)\, \mathrm{d}\Gamma(x, y). \tag{6}$$

The only conceptual change from Monge to Kantorovich is that we no longer insist on a unique destination for each source point. Instead, we optimise over couplings Γ, which can split mass and are therefore much more flexible. In discrete settings, Γ can be viewed as a nonnegative matrix whose row/column sums equal the source/target histograms (see Figure 3). In RL applications, $\mathcal{X}$ and $\mathcal{Y}$ may be action spaces, state spaces, or sets of trajectory features.

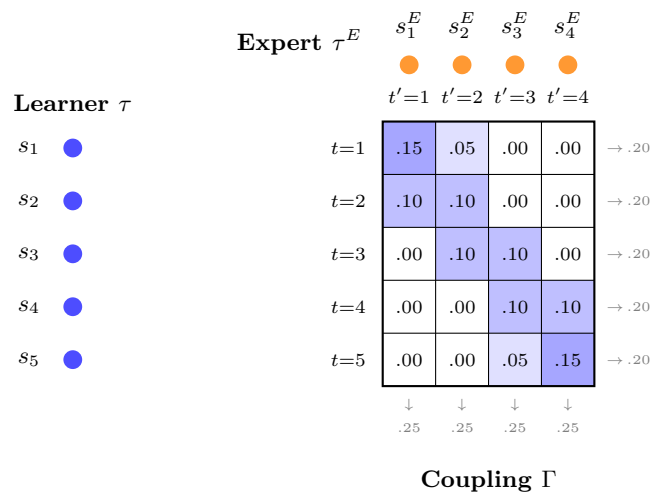

Figure 3: Offline reward labelling via optimal transport (OTR). An optimal coupling Γ^* is computed between empirical state marginals of learner and expert trajectories, treating each as a uniform distribution over its states. This figure shows an example of a coupling, where each row sums to $1/T = 0.20$; each column sums to $1/T_E = 0.25$. The reward at each learner state s_t is the negative transport cost to matched expert states under the optimal plan, $r(s_t) = -\sum_{t'=1}^{T_E} c(s_t, s_{t'}^E)\,\Gamma^*_{t,t'}$. Near-diagonal structure emerges when trajectories are temporally aligned. The coupling in the figure is an example and is not necessarily the optimal coupling.

The associated (general) Kantorovich dual problem is

$$\mathrm{OT}_c(\mu,\nu) = \sup_{u,v}\left\{\int_{\mathcal{X}} u(x)\,\mathrm{d}\mu(x) + \int_{\mathcal{Y}} v(y)\,\mathrm{d}\nu(y) \;:\; u(x)+v(y)\le c(x,y) \text{ for } \mu\text{-a.e. } x \text{ and } \nu\text{-a.e. } y\right\},$$

where u and v are dual potentials. Readers who mainly care about applications can think of the primal problem as "find the best transport plan" and the dual problem as "find the most discriminating test functions consistent with the cost".

Equivalently, one may write

$$\mathrm{OT}_c(\mu,\nu) = \sup_f \left[\int f^c(x)\,\mathrm{d}\mu(x) + \int f(y)\,\mathrm{d}\nu(y)\right]$$

where $f^c(x) = \inf_y\{c(x,y) - f(y)\}$ is the c-transform of f. This single-potential formulation is often more convenient in optimisation and underpins several dual OT solvers used later in the survey.

Rewriting the c-transform yields a maximin formulation:

$$\mathrm{OT}_c(\mu,\nu) = \max_f \min_{F:\,F_\#\mu=\nu} \mathbb{E}_{x\sim\mu}[c(x,F(x)) - f(F(x))] + \mathbb{E}_{y\sim\nu}[f(y)],$$

which is used, for example, in dual formulations of partial OT for offline RL (Section 6.2). For most readers, the important point is not the algebra itself, but that these equivalent forms give different computational handles on the same OT problem.

Wasserstein distances and the W_1 dual form. When $\mathcal{X} = \mathcal{Y}$ is equipped with a ground metric d and $c(x,y) = d(x,y)^p$ for $p \in [1,\infty)$, the p-Wasserstein distance is

$$\mathsf{W}_p(\mu,\nu) = \Big(\inf_{\Gamma\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p\,\mathrm{d}\Gamma(x,y)\Big)^{1/p}.$$

For $p = 1$ and $c(x,y) = \|x-y\|$, Kantorovich–Rubinstein duality yields

$$\mathsf{W}_1(\mu,\nu) = \sup_{\|g\|_{\mathrm{Lip}}\le 1}\left\{\int g\,\mathrm{d}\mu - \int g\,\mathrm{d}\nu\right\}, \tag{7}$$

which underpins critic-based OT estimators (e.g. WGAN-style training) used in several RL objectives. In RL terms, the dual test function g is exactly the "1-Lipschitz critic" that can be trained from samples and then used as a shaping signal (e.g., imitation via occupancy matching). This dual form is one of the main reasons W_1 is so popular in machine learning: it replaces an optimisation over couplings by an optimisation over a critic function that can be parameterised by a neural network.

W_2 structure and transport maps. With quadratic cost on $\mathbb{R}^d$, $c(x, y) = \|x - y\|^2$, the squared 2-Wasserstein distance is

$$\mathsf{W}_2^2(\mu, \nu) = \inf_{\Gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, \mathrm{d}\Gamma(x, y).$$

If μ is absolutely continuous (and mild regularity holds), Brenier's theorem implies that an optimal coupling is induced by a map $T = \nabla\varphi$ where the potential φ is convex and $T_{\#}\mu = \nu$ (Brenier, 1991). This observation motivates *map-based* approximations to W_2 in learning pipelines: one parameterises convex potentials, for example with input-convex architectures, and uses the induced gradient map. In practice, W_2 structure is particularly convenient when distributions are approximately Gaussian or when one wants explicit transport maps.

Entropic regularisation, Sinkhorn iterations, and Sinkhorn divergence. For empirical (often discrete) measures, solving OT_c exactly can be costly. A standard remedy is entropic regularisation (Cuturi, 2013):

$$\mathrm{OT}_\varepsilon(\mu, \nu) = \inf_{\Gamma \in \Pi(\mu,\nu)} \Big\{ \int c(x, y) \, \mathrm{d}\Gamma(x, y) \; + \; \varepsilon \, \mathrm{KL}\big(\Gamma \,\big\|\, \mu \otimes \nu\big) \Big\},$$

where $\mu \otimes \nu$ is the product measure and KL is the Kullback–Leibler divergence. The regularised objective is strictly convex and can be solved efficiently by Sinkhorn (matrix-scaling) iterations. When $c(x, y) = d(x, y)^p$ we write (with an abuse of notation after dropping the p-dependence)

$$\mathsf{W}_\varepsilon^p(\mu, \nu) = \inf_{\Gamma \in \Pi(\mu,\nu)} \Big\{ \int d(x, y)^p \, \mathrm{d}\Gamma(x, y) \; + \; \varepsilon \, \mathrm{KL}\big(\Gamma \,\big\|\, \mu \otimes \nu\big) \Big\}$$

for the entropy-regularised Wasserstein distance. Figure 4 illustrates the effect of entropic regularisation on continuous transport plans, which leads to a smoother coupling. Because W_ε is biased by the entropy term, one often switches to the debiased *Sinkhorn divergence* (Genevay et al., 2018)

$$\mathsf{S}_\varepsilon(\mu, \nu) = \mathsf{W}_\varepsilon(\mu, \nu) - \tfrac{1}{2}\mathsf{W}_\varepsilon(\mu, \mu) - \tfrac{1}{2}W_\varepsilon(\nu, \nu)$$

which satisfies $\mathsf{S}_\varepsilon(\mu, \mu) = 0$ (whereas generally $\mathsf{W}_\varepsilon(\mu, \mu) \neq 0$). This debiasing matters in RL because otherwise the OT loss keeps exerting pressure even when agent and target distributions already match. However, entropy-regularisation without correction is widely used in existing works. A useful way to read this correction is that it subtracts away the artificial self-transport cost created by entropic smoothing. At a high level, entropic regularisation trades exactness for smoothness: instead of a sparse, nearly hard assignment between points, one obtains a softer plan that spreads some mass across multiple destinations. In practice, ε should be chosen on the same scale as typical values of the ground cost c. If it is too small, the plan becomes sharp but Sinkhorn iterations slow down and become numerically fragile. If it is too large, the transport plan is oversmoothed and the task geometry is washed out. Log-domain Sinkhorn is the standard stabilisation trick in the small-ε regime. In RL pipelines, this entropic form is usually the default whenever OT must be computed repeatedly on minibatches, for example in behaviour regularisers, reward labelling, or trust-region penalties.

Heterogeneous spaces: Gromov–Wasserstein. When domains differ, a meaningful cross-space cost $c(x, y)$ may be unavailable (e.g. sim-to-real, different morphologies, or different sensors). Gromov–Wasserstein (GW) (Mémoli, 2011) compares *relations* rather than points: given intra-space dissimilarities $d_\mathcal{X}$ on $\mathcal{X}$ and $d_\mathcal{Y}$ on $\mathcal{Y}$, GW seeks a coupling $\Gamma \in \Pi(\mu, \nu)$ that best matches pairwise distances,

$$\inf_{\Gamma \in \Pi(\mu,\nu)} \iint \big| d_\mathcal{X}(x, x') - d_\mathcal{Y}(y, y') \big|^p \, \mathrm{d}\Gamma(x, y) \, \mathrm{d}\Gamma(x', y').$$

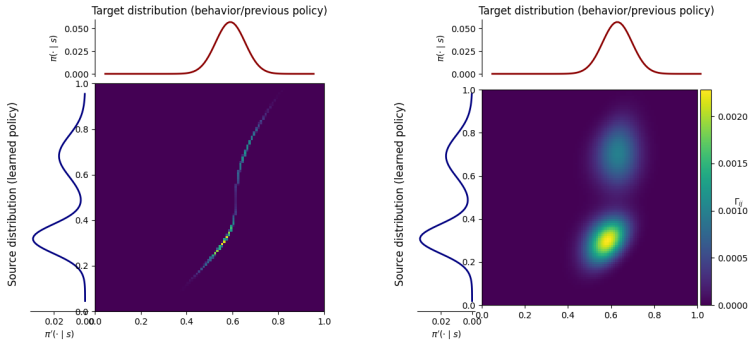

Figure 4: Example of optimal transport between two continuous distributions. The heatmap in the left figure shows the standard OT plan, and the right figure shows the entropy-regularised OT plan ($\varepsilon = 0.01$) which is smoother. Brighter colour indicates more mass is transported.

GW is therefore the natural OT tool when structural alignment matters but a pointwise cross-domain cost is unavailable. In RL, this situation arises most clearly in cross-domain imitation and transfer, where within-domain geometry is trustworthy but pointwise cross-domain comparisons are not. The simplest intuition is that GW asks not "which point matches which point?", but "which neighbourhood structure in one domain resembles which neighbourhood structure in the other?" For example, suppose one agent has three states arranged in a line and another has three visually different states also arranged in a line. A pointwise cost between raw observations may be meaningless, but GW can still recognise that the two systems share the same internal geometry.

Mass mismatch: unbalanced and partial OT. Balanced OT enforces exact marginal constraints, which can be undesirable when demonstrations are partial, contain outliers, or reflect systematic mass mismatch. *Unbalanced OT* relaxes the marginal constraints by penalising deviations of the coupling's marginals from μ and ν (often using KL-type penalties). A common KL-penalised form is

$$\inf_{\Gamma \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \Big\{ \int c \, \mathrm{d}\Gamma + \lambda_1 \, \mathrm{KL}\big((P_{\mathcal{X}})_\# \Gamma \parallel \mu\big) + \lambda_2 \, \mathrm{KL}\big((P_{\mathcal{Y}})_\# \Gamma \parallel \nu\big) \Big\},$$

where $\mathcal{M}_+$ denotes finite nonnegative measures on $\mathcal{X} \times \mathcal{Y}$. As $\lambda_1, \lambda_2 \to \infty$ one recovers the balanced formulation. *Partial OT* instead transports only a prescribed mass $m \leq 1$ by restricting to couplings with total mass m (and sub-marginal constraints):

$$\Pi_m(\mu, \nu) = \Big\{ \Gamma \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y}) : \ (P_{\mathcal{X}})_\# \Gamma \leq \mu, \ (P_{\mathcal{Y}})_\# \Gamma \leq \nu, \ \Gamma(\mathcal{X} \times \mathcal{Y}) = m \Big\},$$

and minimising $\int c \, \mathrm{d}\Gamma$ over $\Pi_m(\mu, \nu)$. These variants are the natural OT tools when the dataset should not be matched in full (e.g. suboptimal demonstrations, incomplete coverage). In RL terms, they allow "some mass to be ignored" (partial) or "mass to be created/removed at a price" (unbalanced), which is often more faithful to real datasets. Figure 5 contrasts the three formulations on a simple example. The distinction is easiest to remember this way: partial OT says "match only the best part of the data," whereas unbalanced OT says "match everything you can, but allow mass mismatch if paying the penalty is cheaper."

Barycentres and interpolation. *p-Wasserstein barycentres* interpolate between distributions by minimising a weighted sum of Wasserstein distances,

$$\eta^\star = \operatorname*{argmin}_{\eta} \sum_k \lambda_k \, \mathsf{W}_p^p(\eta, \mu_k), \qquad \lambda_k \geq 0, \ \sum_k \lambda_k = 1, \tag{8}$$

and appear in curricula (interpolating task distributions) and in some multi-agent consensus formulations. They can be viewed as the OT analogue of "averaging" distributions in a geometry-aware way, and form the basis of displacement interpolation techniques widely used in graphics and imaging (Bonneel et al., 2011).

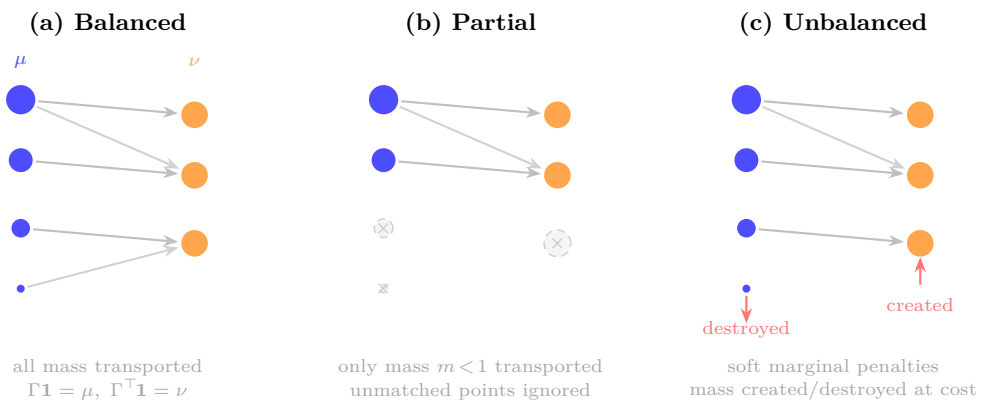

Figure 5: Three OT formulations. (a) Balanced OT enforces exact marginal constraints: all source mass reaches the target. (b) Partial OT transports only a fraction m of the total mass, leaving the rest unmatched ($\times$)—useful when data contain outliers or suboptimal demonstrations (Section 6). (c) Unbalanced OT allows mass creation and destruction at a cost (red arrows), softly penalising marginal deviations rather than enforcing them exactly. Dot sizes indicate relative mass.

Sliced distances and scalable approximations. Several works in the survey use *sliced* (or generalised sliced) Wasserstein distances (Kolouri et al., 2019) as computationally efficient surrogates in high dimensions. The sliced Wasserstein distance compares one-dimensional projections:

$$\mathsf{SW}_p^p(\mu,\nu) = \int_{\theta\in\mathbb{S}^{d-1}} \mathsf{W}_p^p\big((\mathrm{Proj}_\theta)_\#\mu, (\mathrm{Proj}_\theta)_\#\nu\big)\,\mathrm{d}\theta,$$

where $\mathrm{Proj}_\theta(x) = \langle x, \theta\rangle$ is the linear projection on direction θ. It is estimated in practice by Monte Carlo sampling, and generalised variants replace the linear projection by more expressive transforms. The intuition is that 1D Wasserstein distances are cheap with a closed form, and averaging over many slices yields a tractable proxy for high-dimensional transport. Recent work refines this approximation by using sliced generalised Wasserstein geodesics to better capture transport structure beyond random projections (Mahey et al., 2023), and by learning differentiable approximations to generalised sliced transport plans to improve optimisation stability and computational efficiency (Chapel et al., 2025).

Which OT variant when? The survey repeatedly returns to a small set of modelling choices: (i) W_1 *(dual critic):* when supports may be disjoint and one wants a critic-like shaping signal (imitation, some robustness/regularisation); (ii) *entropy-regularised* W_ε*:* when a smooth transport plan, stable gradients and improved computational efficiency is preferred (e.g. in large-scale applications); (iii) *Sinkhorn divergence* S_ε*:* when one wants entropic stability but also a debiased, distance-like quantity; (iv) W_2*:* when closed forms are available with Gaussian distributions or transport maps are desired using a convex potential; (v) *GW:* when cross-domain pointwise costs are ill-posed but relational structure is meaningful (cross-domain imitation/transfer); (vi) *partial/unbalanced OT:* when demonstrations/data are incomplete, contaminated, or systematically mismatched in mass; (vii) *sliced OT:* when dimension makes full OT expensive and projections are adequate; (viii) *OT with windowed cost or masking:* when the temporal structure is crucial.

Conventions for the remainder of the survey. Couplings are denoted Γ and $\Pi(\mu,\nu)$ indicates balanced marginal constraints unless stated otherwise; W_p denotes the p-Wasserstein distance, W_ε the entropy-regularised Wasserstein distance, and S_ε the Sinkhorn divergence. When we do not wish to specify a particular variant of Wasserstein distance we use W. When we speak of a *critic*, we mean a 1-Lipschitz potential in the W_1 dual in Section 3 and value function in the actor-critic setting later; when we speak of a *map*, we mean a W_2-motivated transport map (typically $T = \nabla\varphi$ in Euclidean settings). GW and mass-relaxed variants (unbalanced/partial) are invoked when spaces or masses do not match.

Cost design and numerical scaling. The ground cost $c(x, y)$ (or metric $d(x, y)$ when $c = d^p$) is the main modelling degree of freedom in OT-based RL. In RL applications, the generic points x, y may be *states* ($x = s, y = s'$), *actions* ($x = a, y = a'$), *state–action pairs* ($x = (s, a), y = (s', a')$), or learned *trajectory/behaviour embeddings* (e.g. $x = \Phi(\tau)$). In simple control tasks, Euclidean or cosine distances on states/actions may suffice; in high-dimensional or visual domains, costs are typically computed in a learned feature space f, e.g. $c(s, s') = \|f(s) - f(s')\|$ (or a cosine dissimilarity), and analogously for actions when action geometry is non-trivial. Physics- or goal-aware metrics (task coordinates, kinematic distances, safety margins) should be preferred when available, and heterogeneous units must be normalised so that no single coordinate dominates c.

Numerically, entropic regularisation ε should be scaled to the typical magnitude of c (e.g. the median or a nearby percentile of pairwise costs on representative minibatches), and log-domain Sinkhorn stabilises computation for small ε. For W_1 dual critics, approximate 1-Lipschitzness is enforced (gradient penalty or spectral normalisation); for map-based W_2 approximations, convex parameterisations help ensure valid transports. In later sections, we state explicitly the compared objects, any temporal constraints used when comparing trajectories, the representation used to define c or d, and the numerical scaling/stabilisation choices (e.g. ε and Lipschitz/convexity controls) implemented.

3 Imitation learning and reward inference

In many practical settings the reward is unknown or hard to specify, yet demonstrations of desirable behaviour are available. Imitation learning (IL) seeks a policy from demonstrations, typically without access to rewards. Once imitation is written as occupancy matching, the main design choice becomes the discrepancy between the learner's state–action distribution and the expert's (Figure 6). OT is attractive here because it replaces the geometry-blind divergences used in standard adversarial imitation with a discrepancy that remains informative under weak overlap and can incorporate task structure through the ground cost. In practical terms, OT helps answer the question "how far is the learner from the expert, and in what sense?" In this section we start from the occupancy-matching view, then show how that OT signal appears in online reward shaping, offline pseudo-reward labelling, temporally aware matching, and cross-domain alignment.

3.1 Imitation as occupancy matching

It is useful to introduce three main categories of IL because OT will enter naturally only after this setup is clear.

Behaviour cloning (BC) treats imitation as supervised learning on demonstration pairs (s, a), fitting $\pi(\cdot \mid s)$ to the expert's actions. It is simple and effective in-distribution but brittle under covariate shift: small action errors at test time move the agent into states that are rare or absent in the dataset, where the learned conditional $\pi(\cdot \mid s)$ is poorly specified, and errors compound over time. Intuitively, BC asks the right question one step at a time, but has no mechanism for correcting the distribution shift that those one-step mistakes create.

Inverse reinforcement learning (IRL) instead posits an unknown *cost* function $\ell : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ that explains expert behaviour and then seeks a policy that minimises this cost (we use ℓ to avoid clashing with the OT ground cost $c(x, y)$). A key object is the normalised discounted occupancy measure $\rho_\pi(s, a)$ (Section 2), which converts expectations under the trajectory law into expectations under a single distribution on $\mathcal{S} \times \mathcal{A}$. This is the step where the imitation problem becomes distributional: instead of asking whether two policies take the same action in the same state, we ask whether they induce the same long-run visitation pattern. In maximum-entropy RL/IRL one adds the discounted *causal entropy* (Bloem & Bambos, 2014)

$$H(\pi) \;=\; \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty} \gamma^t \big(-\log \pi(a_t \mid s_t)\big)\Big] \;=\; \frac{1}{1-\gamma}\, \mathbb{E}_{(s,a)\sim\rho_\pi}\big[-\log \pi(a \mid s)\big],$$

which discourages overly deterministic policies and yields a well-posed optimisation problem.

A regularised maximum-entropy IRL problem can be written (Ho & Ermon, 2016) as

$$\mathrm{IRL}_\psi(\pi_E) = \operatorname*{argmax}_\ell \Big\{ -\psi(\ell) + \min_\pi \big[-H(\pi) + \mathbb{E}_{(s,a)\sim\rho_\pi}[\ell(s,a)] \big] - \mathbb{E}_{(s,a)\sim\rho_{\pi_E}}[\ell(s,a)] \Big\},$$

with a convex regulariser ψ on costs. The inner minimisation is the entropy-regularised control problem for a fixed ℓ:

$$\pi^\star(\ell) \in \operatorname*{argmin}_\pi \Big\{ -H(\pi) + \mathbb{E}_{(s,a)\sim\rho_\pi}[\ell(s,a)] \Big\}.$$

The standard max-ent IRL pipeline first finds a cost $\ell^\star$ via the outer maximisation and then outputs $\pi^\star(\ell^\star)$. Eliminating ℓ by convex duality yields the equivalent occupancy-matching form

$$\pi^\star \in \operatorname*{argmin}_\pi \Big\{ -H(\pi) + \psi^*\big(\rho_\pi - \rho_{\pi_E}\big) \Big\}, \tag{9}$$

which makes explicit that imitation reduces to matching *occupancy measures* and that the choice of discrepancy ψ^* (the convex conjugate of ψ) is central. This is exactly the point where OT can be inserted: once imitation is "matching these two distributions," the next question is *which* discrepancy should be used to define the match.

Adversarial imitation learning can be viewed as an algorithmic instantiation of this occupancy-matching objective: it chooses a variational form for ψ^* and optimises it using a learned discriminator/critic from samples of expert and learner behaviour. GAIL is the canonical example, corresponding to an f-divergence choice (JS) for ψ^* implemented via an adversarial discriminator (Goodfellow et al., 2014; Ho & Ermon, 2016).

3.2 Why optimal transport helps

The occupancy view makes clear why OT is useful in imitation learning. The learner must bring ρ_π close to ρ_{π_E}, but early in training these distributions often overlap very little. As discussed in Section 2.2, f-divergences become weak or unstable in exactly this regime, whereas OT still measures how much probability mass must move, and how far, under a task-dependent ground cost. That ground cost may live on $\mathcal{S} \times \mathcal{A}$, on $\mathcal{S}$ alone in learning-from-observation settings, or in a learned feature space. In each case it determines what the algorithm treats as a meaningful notion of behavioural similarity. Put differently, OT does not just say "the learner is wrong"; it says *how* the learner is wrong in the geometry of the task.

The rest of the section follows two questions: *which distributions are being compared*, and *how does the OT signal enter learning*? This leads to the following recurring patterns. A reader can think of them as different places where OT can sit in the pipeline: inside an online reward signal, inside an offline relabelling step, inside the coupling constraints, or inside the representation used for matching.

- *Occupancy matching with OT discrepancies.* Replace the JS/KL discrepancy in adversarial IL by a Wasserstein objective (typically W_1 via a 1-Lipschitz critic) or by a primal entropy-regularised objective, yielding geometry-aware shaping rewards.
- *Reward labelling from couplings (offline).* Compute a transport plan between empirical measures built from learner and expert trajectories and convert the resulting coupling into per-state (or per-transition) pseudo-rewards, enabling standard offline RL without an online min–max loop for reward learning in standard IRL. With multiple expert demonstrations, multi-marginal/barycentric aggregation can avoid the noise induced by naive concatenation (Sebag et al., 2023).
- *Temporal and intent-aware matching.* Modify admissible couplings (e.g. banded/time-window constraints) and/or modify the ground cost (e.g. intent/successor-feature embeddings) to reduce spurious matches caused by treating trajectories as unordered bags of states.
- *Cross-domain matching.* When a pointwise cross-domain ground cost is ill-posed, use relational objectives such as Gromov–Wasserstein that match within-domain geometry rather than raw observations.
- *Auxiliary OT scores.* Reuse OT-derived costs as filtering/reweighting signals before downstream offline RL when data come from multiple domains (Section 7.6).

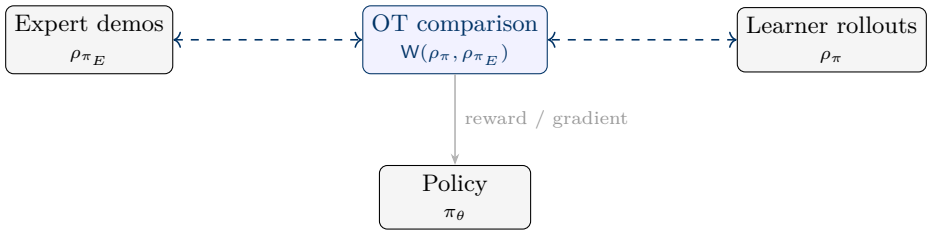

Figure 6: OT in imitation learning and reward inference (Section 3). OT compares learner and expert behaviour distributions via dual critics or primal couplings, yielding shaping rewards or pseudo-reward labels (Sections 3.3–3.4). Temporal/intent-aware extensions modify the coupling or ground cost (Section 3.5); cross-domain settings use GW (Section 3.6).

3.3 Reward shaping in the online adversarial setting

The max-ent IRL derivation in Section 3.1 poses imitation as occupancy matching (9), where ψ^* measures the gap between learner and expert occupancies (Ho & Ermon, 2016). Adversarial imitation learning turns this variationally into a learned critic and uses the critic output as a shaping reward for policy optimisation. GAIL is the canonical example: it chooses ψ^* so that the induced discrepancy is an f-divergence, specifically JS, implemented with an adversarial discriminator (Ho & Ermon, 2016). OT-based variants keep this overall template but replace the underlying discrepancy by a Wasserstein objective, or by a tractable OT-inspired approximation, so that the reward signal remains informative under weak overlap and can reflect task geometry. The practical picture is simple: OT is used to turn the different behaviour between expert and learner into a dense reward-like signal that the policy can improve against.

A useful way to organise the OT-based occupancy-matching literature is by *how the Wasserstein-type discrepancy is computed*.

Dual (critic-based) W_1: Kantorovich potentials as shaping rewards. When the discrepancy is W_1 on $\mathcal{S} \times \mathcal{A}$, Kantorovich–Rubinstein duality (Section 2.2) expresses $\mathsf{W}_1(\rho_\pi, \rho_{\pi_E})$ as a supremum over 1-Lipschitz critics f, suggesting a WGAN-style scheme (Arjovsky et al., 2017): train f to maximise the dual gap between expert and learner occupancies, then use $\tilde{r}(s,a) = f(s,a)$ as a shaping reward. Wasserstein Adversarial Imitation Learning (WAIL) instantiates this idea and approximates the W_1 dual by replacing the hard 1-Lipschitz constraint with a penalty $\Omega_{d,\varepsilon}$ (Xiao et al., 2019):

$$\mathsf{W}_1(\rho_\pi, \rho_{\pi_E}) = \sup_{f:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \Big\{ \mathbb{E}_{y\sim\rho_{\pi_E}}[f(y)] - \mathbb{E}_{x\sim\rho_\pi}[f(x)] + \mathbb{E}_{(x,y)\sim\rho_\pi\times\rho_{\pi_E}}\big[\Omega_{d,\varepsilon}(f,x,y)\big] \Big\}.$$

The practical requirement in dual-W_1 approaches is stable enforcement (or relaxation) of 1-Lipschitzness, commonly via weight clipping or gradient penalties (Arjovsky et al., 2017; Gulrajani et al., 2017). Follow-up works explore alternative design choices and application settings (e.g. different shaping rewards or state-only matching); Zhang et al. (2020); Wu et al. (2023) are representative examples.

Primal entropic OT (Sinkhorn): coupling-derived shaping rewards. An alternative is to compute an OT plan directly in the primal using entropic regularisation and Sinkhorn iterations (Cuturi, 2013; Genevay et al., 2018). SIL (Papagiannis & Li, 2022) formulates imitation as minimising an entropy-regularised Wasserstein distance W_ε between expert and learner occupancy measures and uses the resulting coupling to derive a dense shaping reward. The main advantage of this route is conceptual as well as computational: one gets not just a scalar discrepancy, but an explicit soft matching between learner and expert samples. A key modelling choice is the ground cost: SIL defines a cosine dissimilarity in an adversarially learned feature space, so the cost matrix is both geometry-aware and discriminative. The feature-space parameters are trained by maximising W_ε between learner and expert occupancies, yielding a minimax objective analogous to GAIL but with a Sinkhorn-based discrepancy (Papagiannis & Li, 2022). Relative to dual-W_1 critics, the entropic primal objective provides smooth, minibatch-friendly estimates of the discrepancy and an explicit transport plan; in exchange, performance depends on scaling the entropic parameter and the feature-space cost appropriately.

Computational approximation: greedy primal couplings. Some methods try to reduce the cost of repeatedly solving OT subproblems.

PWIL avoids adversarial training and approximates the primal OT coupling by a greedy strategy that yields an upper bound on the true Wasserstein cost (Dadashi et al., 2021a). At each step, the agent's current state–action is matched to nearby expert samples subject to remaining "mass", producing a non-stationary cost $c^g_{t,\pi}$ and a shaping reward $\tilde{r}_{t,\pi} = f(c^g_{t,\pi})$ for each state-action pair with a monotone decreasing function f (Dadashi et al., 2021a). This preserves the geometric "reward-from-transport" intuition while avoiding a min–max loop, at the expense of a history-dependent reward induced by the evolving greedy coupling.

All three approaches pursue the same goal, namely to turn transport-based distribution matching into a usable reward signal, but they do so in different computational regimes. Dual-W_1 critics are conceptually simple but live or die by Lipschitz control. Sinkhorn-primal methods give an explicit coupling and behave well on minibatches, but introduce the usual sensitivity to the entropic parameter. Greedy couplings such as PWIL avoid adversarial training altogether, at the price of a history-dependent reward.

3.4 Offline reward labelling

Instead of learning a shaping signal online through adversarial occupancy matching, one can compute pseudo-rewards *offline*. The idea is to align trajectories in an unlabeled dataset with expert demonstrations using OT, then read per-time-step labels from the resulting coupling and pass the relabelled dataset to a standard offline RL algorithm. The important shift is that OT is now solved between *empirical measures* built from finite trajectories, rather than between discounted occupancies (Luo et al., 2023). In effect, the transport plan is being used as a soft correspondence map: it says which parts of a learner trajectory resemble which parts of an expert trajectory.

Concretely, consider an offline dataset $\mathcal{D}$ consisting of transitions (s, a, s') (with rewards unavailable or ignored) and a set of expert demonstrations. For a single episode in $\mathcal{D}$ we write

$$\tau = \big((s_t, a_t, s_{t+1})\big)_{t=0}^{T-1},$$

and denote by $(s_0, \ldots, s_{T-1})$ the associated state sequence on which the OT problem is constructed. An expert demonstration may be available either as a full trajectory (including actions) or as an observation-only sequence; in both cases we denote the expert state sequence by $(s^E_0, \ldots, s^E_{T_E-1})$.

OTR (Luo et al., 2023) constructs empirical state measures

$$\hat{p}_\pi = \frac{1}{T}\sum_{t=0}^{T-1} \delta_{s_t}, \qquad \hat{p}_E = \frac{1}{T_E}\sum_{t'=0}^{T_E-1} \delta_{s^E_{t'}},$$

and, given a ground cost $c : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$ (e.g. cosine distance on raw state vectors), computes an optimal coupling

$$\Gamma^* = \operatorname*{argmin}_{\Gamma \in \Pi(\hat{p}_\pi, \hat{p}_E)} \sum_{t=0}^{T-1}\sum_{t'=0}^{T_E-1} c(s_t, s^E_{t'})\, \Gamma_{t,t'}, \qquad \Pi(\hat{p}_\pi, \hat{p}_E) = \big\{\Gamma \in \mathbb{R}^{T\times T_E} :\ \Gamma\mathbf{1} = \tfrac{1}{T}\mathbf{1},\ \Gamma^\top\mathbf{1} = \tfrac{1}{T_E}\mathbf{1}\big\}.$$

A per-step pseudo-reward is then assigned by aggregating transport costs under Γ^*:

$$\tilde{r}_{\mathrm{OT}}(s_t) \ = \ -\sum_{t'=0}^{T_E-1} c\big(s_t, s^E_{t'}\big)\, \Gamma^*_{t,t'}.$$

Computing Γ^* exactly is a linear program; in practice OTR solves an entropy-regularised OT problem and uses Sinkhorn iterations for efficiency (Luo et al., 2023; Cuturi, 2013). After squashing through an exponential function (e.g. $s(r) = 5\exp(Tr)$), rewards are rescaled and independent of the horizon. These labels are attached to the offline transitions at the corresponding time steps (e.g. $(s_t, a_t, s_{t+1}) \mapsto s(\tilde{r}_{\mathrm{OT}}(s_t))$) and then used by any downstream RL method (Haldar et al., 2023; Wang et al., 2023; Zare et al., 2023).

Learning from observation. When expert actions are unavailable (imitation learning from observation, ILfO), Chang et al. (2023) argue that state-only matching can be problematic in some environments depending on the state representation and dynamics, and treat trajectories as bags of *transitions* rather than states, defining empirical measures $\hat{p}_\pi^{\rightarrow} = \frac{1}{T-1}\sum_t \delta_{(s_t,s_{t+1})}$ and $\hat{p}_E^{\rightarrow} = \frac{1}{T_E-1}\sum_{t'} \delta_{(s^E_{t'},s^E_{t'+1})}$ on $\mathcal{S}\times\mathcal{S}$ and solving OT with a ground metric d on transition pairs to yield per-transition pseudo-rewards $\tilde{r}_{\mathrm{OT}}(s_t,s_{t+1}) = -\sum_{t'} d\big((s_t,s_{t+1}),(s^E_{t'},s^E_{t'+1})\big)\,\Gamma^{\rightarrow *}_{t,t'}$. They also show that under a monotonic-improvement assumption, reusing pseudo-rewards from previously collected trajectories yields an upper bound on the current Wasserstein objective, enabling systematic off-policy reuse.

In offline learning-from-observations (LfO) where interaction is disallowed, DICE (Distribution Correction Estimation) methods (Nachum et al., 2019) reformulate occupancy matching via a change of variables that avoids explicit policy evaluation, instead learning density-ratio corrections directly from data. They seek a policy whose *state* occupancy $\rho_\pi(s) = \sum_a \rho_\pi(s,a)$ matches an empirical expert state occupancy estimated from expert state-only trajectories. PW-DICE (Yan et al., 2024) proposes to minimise the Kantorovich primal (6) on state occupancies, $\mathsf{W}_1(\rho_\pi(s),\rho_{\pi_E}(s))$, jointly over a feasible occupancy induced by π and a coupling Γ. Feasibility is enforced via the Bellman-flow constraint (4) on ρ_π, together with $\rho_\pi(s) = \sum_a \rho_\pi(s,a)$ and nonnegativity. The main reason PW-DICE adopts the primal view is flexibility of the ground cost. Dual-W_1 methods usually enforce Lipschitzness through Euclidean gradient penalties, which quietly ties the geometry to Euclidean structure. The primal formulation instead allows arbitrary costs $c(s,s')$, including learned reachability-aware metrics (Yan et al., 2024). On the optimisation side, PW-DICE adds pessimistic regularisers, implemented with KL-type penalties, to obtain a smooth convex dual objective, then recovers the policy by weighted behaviour cloning. The learned dual variables play a role analogous to rewards in offline reward-labelling methods from Section 3.4. The same formulation also shows that, for particular hyperparameter choices, familiar f-divergence DICE variants appear as special cases, which gives a clean unifying view of Wasserstein and f-divergence occupancy matching in offline LfO (Yan et al., 2024).

Two limitations motivate the extensions in Section 3.5. First, state-marginal transport ignores order and can couple reversed or speed-mismatched rollouts. Second, the pseudo-rewards are context-dependent: $\tilde{r}_{\mathrm{OT}}(s_t)$ depends on the whole trajectory through Γ^*, so the same state may receive different labels across episodes. Put more intuitively, a bag-of-states OT problem can say that two trajectories contain similar ingredients without checking whether those ingredients appear in the same order. Temporal constraints (e.g. banded couplings and context-window costs) and intent-aware costs mitigate these issues by biasing the coupling toward locally consistent alignments and control-relevant geometry (Fu et al., 2024; Bobrin et al., 2024).

Multiple expert demonstrations: concatenation vs. multi-marginal/barycentric aggregation. The construction above is often presented for a single demonstration, but practical IL usually offers a *set* of expert trajectories $\{\tau_E^{(p)}\}_{p=1}^P$, often with clear multi-modality. A common heuristic is simply to pool all expert states into one empirical measure and then subsample. When demonstrations differ substantially, however, this can wash out modes and produce noisy pseudo-rewards. Sebag et al. (2023) address this problem directly in SMMOTIL by combining demonstrations in OT geometry, via sliced multi-marginal OT distances, rather than by concatenation. Their method keeps the reward-from-transport viewpoint of PWIL but changes how expert information is aggregated. Sebag et al. (2023) implement this idea through *sliced* approximations, which reduce the complexity of transport computations by averaging one-dimensional Wasserstein distance over random projections and admit closed-form 1D solvers for discrete measures. Concretely, their multi-marginal variant computes a sliced multi-marginal Monge–Wasserstein discrepancy between the agent rollout and all expert rollouts jointly, which is theoretically equivalent to comparing the agent rollout to the sliced-Wasserstein barycentre of experts with W_2. Empirically, this reduces instability and improves performance when expert demonstrations are diverse.

3.5 Temporal and intent-aware extensions

The trajectory-level construction in the previous subsection treats rollouts as bags of states. This simplifies transport but discards ordering, so the OT plan can align temporally distant (or even reversed) parts of learner and expert trajectories (Figure 7). The resulting pseudo-rewards can therefore be noisy and overly dependent on the global trajectory-level alignment. Two complementary fixes have been proposed: (i) enforce

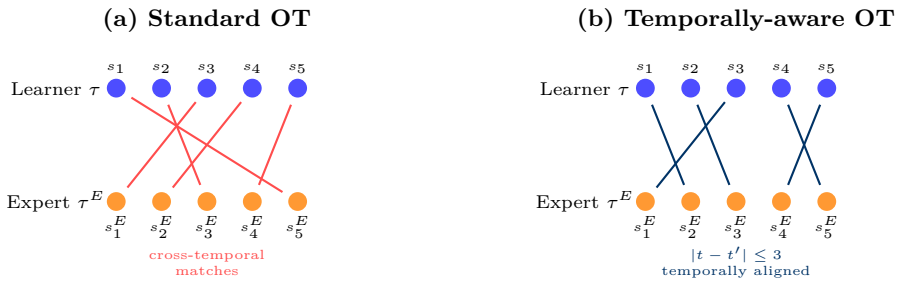

Figure 7: Standard vs. temporally-aware OT matching between learner and expert trajectories. (a) Without temporal constraints, the coupling can align temporally distant states, producing crossing matches that yield noisy pseudo-rewards. (b) With a band constraint $|t - t'| \leq 3$ (or context-window costs), matches are near-diagonal, reducing spurious alignments while tolerating modest speed variation.

approximate time alignment by restricting the coupling to near-diagonal time indices and using windowed (context) costs, and (ii) define the ground cost in embeddings that reflect short-horizon dynamics ("intent"), so matches depend on where the agent is going rather than only what it looks like now.

Temporal masking and context smoothing. A first strategy constrains the coupling to respect approximate time alignment. TemporalOT (Fu et al., 2024) proposes two coupled modifications: (i) a *context* cost that compares short windows rather than single frames, and (ii) a *band* constraint that restricts admissible couplings to near-diagonal indices.

Let $(o_0, \ldots, o_{T-1})$ and $(o_0^E, \ldots, o_{T-1}^E)$ be learner and expert observation sequences of equal length, and let $f(o)$ be a fixed visual encoder (TemporalOT targets image-based tasks in partially observable settings; in fully observable MDPs, observations reduce to states). For a context length k_c, define the group-wise cosine dissimilarity cost (Fu et al., 2024)

$$\hat{c}(o_i, o_j^E) = \frac{1}{k_c} \sum_{h=0}^{k_c-1} \Big(1 - \frac{\langle f(o_{i+h}),\, f(o_{j+h}^E)\rangle}{\|f(o_{i+h})\|_2\, \|f(o_{j+h}^E)\|_2}\Big),$$

with indices clipped when $i + h$ or $j + h$ exceeds the horizon. For a band width k_m, define a binary mask

$$M(i, j) = \mathbb{1}\big(|i - j| \leq k_m\big),$$

and restrict the coupling between uniform empirical marginals $\hat{p}_\pi = \frac{1}{T}\sum_{i=0}^{T-1} \delta_{o_i}$ and $\hat{p}_E = \frac{1}{T}\sum_{j=0}^{T-1} \delta_{o_j^E}$ to the near-diagonal set

$$\Pi_M(\hat{p}_\pi, \hat{p}_E) = \Big\{\Gamma \in \Pi(\hat{p}_\pi, \hat{p}_E) :\ \Gamma_{ij} = 0 \text{ whenever } M(i, j) = 0\Big\}.$$

TemporalOT then solves an entropic OT problem on this restricted set,

$$\Gamma^* = \underset{\Gamma \in \Pi_M(\hat{p}_\pi, \hat{p}_E)}{\operatorname{argmin}} \Big\{\sum_{i,j} \Gamma_{ij} \hat{c}(o_i, o_j^E) - \varepsilon\, \mathcal{H}(\Gamma)\Big\},$$

which is equivalent to running Sinkhorn with forbidden entries assigned an infinite (or very large) cost. The resulting Γ^* can be converted into a dense pseudo-reward by aggregating costs under the coupling, as in Section 3.4; empirically, intermediate k_m and k_c reduce spurious matches while tolerating modest speed variation (Fu et al., 2024). The mask gives a temporally informed proxy reward by discouraging cross-time matches, which is not a full dynamic OT formulation.

Intent representations and short-horizon dynamics. The banded coupling above addresses *time-index* ambiguity, but it does not resolve a second failure mode: even near-diagonal matches can be spurious if the ground cost compares only instantaneous observations. A complementary strategy is therefore to encode

intent directly in the representation used for the ground cost, so that similarity reflects short-horizon evolution ("where the agent is going") rather than appearance alone.

AILOT (Bobrin et al., 2024) instantiates this by learning an intent embedding $\varphi : \mathcal{S} \to \mathbb{R}^d$ and defining the OT cost on *paired* intent states separated by a fixed lookahead:

$$c(s_i, s_{\min(i+k,T-1)}, s_j^E, s^E_{\min(j+k,T_E-1)}) = \|\varphi(s_i) - \varphi(s_j^E)\|_2^2 + \|\varphi(s_{\min(i+k,T-1)}) - \varphi(s^E_{\min(j+k,T_E-1)})\|_2^2,$$

where k probes near-future evolution. Given the entropic OT plan Γ^* between the two trajectories, AILOT defines a dense intrinsic pseudo-reward by exponentially scaling the transport cost aggregated under Γ^*,

$$\tilde{r}(s_i) = \alpha \exp\Big(- \kappa T \sum_{j=j_1}^{T_E-1} \Gamma^*_{ij}\, c(s_i, s_{\min(i+k,T-1)}, s_j^E, s^E_{\min(j+k,T_E-1)})\Big),$$

$$j_1 = \operatorname*{argmin}_j c(s_0, s_{\min(k,T-1)}, s_j^E, s^E_{\min(j+k,T_E-1)}),$$

where restricting to the "tail" $j \geq j_1$ anchors alignment to the closest expert prefix for the given agent start state. Intuitively, the paired-window cost rewards states that not only look similar now but are poised to evolve similarly over the next few steps, improving robustness to nuisance variation and one-step ambiguities. In control terms, this aligns trajectories by *intent* rather than appearance.

3.6 Cross-domain imitation with Gromov–Wasserstein

The OT machinery used for imitation can also serve cross-domain settings where expert and learner observations live in different spaces, so that any pointwise cross-domain cost is ill-posed. OT-based data filtering for downstream offline RL (e.g. Lyu et al. 2025) is discussed in Section 7.6.

Relational matching via GW. The trajectory transport constructions above assume a shared observation space and a ground cost $c(s, s')$ that is meaningful across expert and learner domains. In many applications this assumption fails (different sensors, viewpoints, or embodiments), so any pointwise cross-domain cost is ill-posed or dominated by nuisance variation.

The Gromov–Wasserstein (GW) distance (Section 2.2) addresses this by matching *relations* rather than points, seeking a coupling Γ that preserves intra-domain pairwise distances. GWIL applies GW to compare expert and learner state–action occupancy measures living in different spaces, and converts the resulting coupling into a pseudo-reward for policy optimisation (Fickinger et al., 2022). In practice, GW objectives are computationally heavier than standard Wasserstein distances; entropic regularisation and associated solvers provide scalable approximations and are widely used in GW implementations (Peyré et al., 2016), building on regularised standard OT (Ferradans et al., 2014). Keypoint-guided OT (Gu et al., 2022), which leverages partial correspondences between source and target, offers an alternative route when some cross-domain anchor pairs are available.

3.7 Summary

Imitation learning can be expressed as matching *policy-induced distributions*, most commonly discounted occupancies ρ_π (learner) and ρ_{π_E} (expert), or empirical trajectory measures in state/observation space. OT is useful in this setting because it remains informative under weak overlap: it measures how far mass must move under a task-dependent ground cost rather than comparing densities pointwise.

Across the methods reviewed in this section, OT enters the learning pipeline in several distinct roles. (i) *Online occupancy matching:* adversarial methods replace the JS/KL-style discriminator with a Wasserstein objective (typically dual W_1 with a 1-Lipschitz critic) or a primal entropic OT/Sinkhorn objective, producing geometry-aware *shaping* rewards. (ii) *Offline reward labelling:* coupling-based methods solve OT between empirical measures built from learner and expert trajectories (states or transitions) and convert the optimal plan Γ^* into per-step *pseudo-reward* labels, enabling standard offline RL without an online min-max loop. (iii) *Temporal/intent structure:* to avoid spurious alignments caused by treating trajectories as unordered

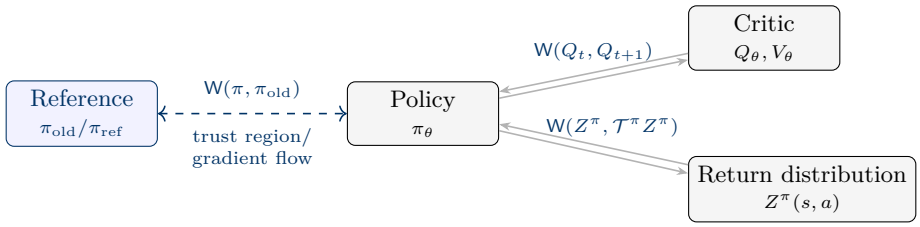

Figure 8: OT in policy optimisation and regularisation (Section 4). Wasserstein trust regions constrain per-state action distributions or trajectory embeddings to stay close to a reference policy (Sections 4.2–4.4). Wasserstein-trained critics use OT losses on return or Q-value distributions (Section 4.5). The blue dashed bidirectional arrow connects two distributions compared by OT, while grey solid arrows represent connections in RL algorithm.

bags of states, extensions either restrict admissible couplings (e.g. near-diagonal bands and context windows) or redefine the ground cost in intent/dynamics-aware embeddings. (iv) *Cross-domain imitation:* when a pointwise cross-domain cost is ill-posed, Gromov–Wasserstein provides a well-motivated alternative by matching relational structure under intra-domain metrics. (v) *Auxiliary OT scores:* OT-derived costs can be reused as filtering/reweighting signals to select relevant data before downstream offline RL. The main failure mode is consistent across these roles: a poor ground cost or weak temporal alignment can turn a transport plan into misleading rewards.

Two further themes connect to later sections. Dynamic/path-space transports (Section 3.5) offer a direct route to time-consistent alignment but remain computationally heavy at long horizons. OT also quantifies IRL reward ambiguity via Wasserstein dispersion (Section 7.5), and OT-derived scores can filter data before downstream offline RL (Section 7.6). In practice, the dominant failure modes are representational—poorly specified ground costs mis-shape behaviour and numerically unstable Lipschitz enforcement or overly diffuse entropic regularisation.

4 Policy optimisation and regularisation

Policy-gradient methods update a parameterised policy by ascending an estimate of the return gradient, and their stability depends heavily on what counts as a "small" policy change. Classical algorithms answer that question with density-ratio geometry: TRPO uses a KL-based trust region, while PPO clips likelihood ratios. OT offers a different answer. Instead of comparing policies only through likelihood ratios, it compares the cost of moving probability mass between their induced action or behaviour distributions. The intuition is that not all policy changes of the same probability size are equally meaningful: shifting mass from one action to a very similar one should count as a smaller change than shifting it to a qualitatively different action. In this section we follow that idea through four settings: trust regions on action distributions, regularisation in latent behaviour spaces, Wasserstein gradient-flow and natural-gradient updates, and critic training based on Wasserstein losses (Figure 8).

4.1 Classical policy optimisation

We begin with the standard policy-optimisation setup, because the OT variants are easiest to understand as direct modifications of familiar trust-region and actor–critic updates. The goal of this subsection is not to re-derive policy gradients in full, but to isolate the place where "geometry" enters the update.

Let π be the current policy and let ρ_π denote the discounted state-action visitation measure induced by π. The control objective (5), recalled from Section 2, expresses $J(\pi)$ as an expectation under the discounted occupancy measure ρ_π. The prefactor $\frac{1}{1-\gamma}$ is constant and can be suppressed when discussing optimisation templates. In policy-gradient methods, one considers a parameterised family $\{\pi_\theta\}$ and updates θ in the direction of the gradient $\nabla_\theta J(\pi_\theta)$. The policy gradient theorem shows that this gradient can be expressed in terms of an *advantage function* A^π as

$$\nabla_\theta J(\pi_\theta) \;=\; \frac{1}{1-\gamma}\mathbb{E}_{s\sim\rho_{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}\big[\nabla_\theta \log \pi_\theta(a \mid s)\, A^{\pi_\theta}(s,a)\big],$$

where $A^\pi(s,a)$ measures how much better action a is, in state s, than the policy's default behaviour in that state (Sutton & Barto, 1998). In practice one replaces A^{π_θ} by an estimate $\hat{A}$ obtained from rollouts and takes stochastic gradient steps.

It is helpful to introduce a candidate next policy π' and ask how much expected advantage it gains under the *state* distribution of the current policy π. This leads to the surrogate objective

$$\mathbb{E}_{s\sim\rho_\pi,\, a\sim\pi'(\cdot|s)}\big[A^\pi(s,a)\big], \tag{10}$$

where A^π is the advantage under the current policy π. The expression (10) keeps the state distribution fixed at what π currently visits, and chooses π' so that it places more probability mass on actions with positive advantage and less on actions with negative advantage. This is why it is called a *surrogate*: it is a local stand-in for the true return objective, using information that is reliable near the current policy. One can show that, for sufficiently small changes in the policy, maximising (10) provides a first-order approximation to improving $J(\pi')$ itself (Schulman et al., 2015).

If (10) is maximised without any restriction, the resulting policy π' can differ arbitrarily from π after a single update. In practice this often leads to instability: the advantage estimates are only reliable near the behaviour that generated the data, and large updates can push the agent into regions where those estimates are inaccurate. Trust-region methods address this by explicitly requiring that, in each state, the new action distribution $\pi'(\cdot \mid s)$ remains close to the old one $\pi(\cdot \mid s)$. A trust region is therefore best thought of as a local safety mechanism: improve the policy, but only by a step whose size is small enough that the local approximation still makes sense.

Trust Region Policy Optimisation (TRPO) (Schulman et al., 2015) enforces this by constraining the average Kullback–Leibler (KL) divergence between the two policies,

$$\mathbb{E}_{s\sim\rho_\nu^\pi}\Big[\mathrm{KL}\big(\pi(\cdot \mid s)\,\|\,\pi'(\cdot \mid s)\big)\Big] \;\le\; \delta,$$

for some step-size parameter $\delta > 0$, where ρ_ν^π denotes the discounted state-visitation distribution induced by policy π from initial-state distribution ν. Under standard assumptions, maximising (10) subject to this constraint yields policy updates that improve performance monotonically and do not move too far in a single step. Proximal Policy Optimisation (PPO) (Schulman et al., 2017) takes a more heuristic approach: it modifies the objective so that, for each state and action, the contribution to the gradient is clipped when the ratio $\pi'(\cdot \mid s)/\pi(\cdot \mid s)$ becomes too large or too small. This clipping effectively limits the change in log-probabilities and plays a role similar to a KL-based trust region. Schulman et al. (2017) also propose the KL penalty with adaptive coefficient as an alternative.

In actor-critic methods, a learned critic is used to approximate the value functions that appear in the advantage. With a parametric value function V_{θ_v} as critic, it is trained with the lost function

$$\mathcal{L}(\theta_v) = \mathbb{E}_\pi[(y_t - V_{\theta_v}(s_t))^2]$$

to predict the temporal-difference (TD) target

$$y_t = r_t + \gamma\, V_{\bar{\theta}_v}(s_{t+1}),$$

according to the Bellman equation (1), where $V_{\bar{\theta}_v}$ is the target with parameters updated slowly to enhance the learning stability. Advantage estimates $\hat{A}_t$ are then constructed, for example via generalised advantage estimation. The policy gradient is then approximated by replacing $A^\pi(s,a)$ with $\hat{A}_t$ in the surrogate objective (10), so that the actor update is driven directly by the critic's outputs. In this setting, stability depends both on the size of the policy step and on how smoothly the critic and its targets evolve.

In all of these algorithms, the notion of a "small" policy update is already present; OT simply changes the geometry used to measure it. TRPO makes the notion explicit with an average KL constraint, while PPO approximates it through clipped likelihood ratios. The remainder of the section keeps this basic optimisation template but replaces KL-style geometry by optimal transport on action distributions, behavioural embeddings, or value distributions.

4.2 Wasserstein trust regions for policies

The most direct OT analogue of TRPO is to replace the KL trust region by a Wasserstein trust region on per-state action distributions.

Let $c : \mathcal{A} \times \mathcal{A} \to \mathbb{R}_{\geq 0}$ be a ground cost on the action space, and let $\mathsf{W}(\cdot, \cdot)$ denote the induced Wasserstein or entropy-regularised Wasserstein distance between probability measures on $\mathcal{A}$ (Cuturi, 2013; Genevay et al., 2018). Recall from Section 2 that $\mathsf{W}(\mu, \nu)$ measures the minimal cost of transporting the mass of μ onto the mass of ν when moving a unit of mass from a to a' incurs cost $c(a, a')$. For a fixed state s, the discrepancy between $\pi(\cdot \mid s)$ and $\pi'(\cdot \mid s)$ can therefore be measured by $\mathsf{W}\big(\pi(\cdot \mid s), \pi'(\cdot \mid s)\big)$ and used in place of KL in the trust-region constraint.

The analogue of (10) with an OT trust region is

$$\max_{\pi'} \; \mathbb{E}_{s\sim\rho^{\pi}_{\nu},\, a\sim\pi'(\cdot|s)}[A^{\pi}(s,a)] \quad \text{s.t.} \quad \mathbb{E}_{s\sim\rho^{\pi}_{\nu}}\, \mathsf{W}(\pi'(\cdot \mid s),\, \pi(\cdot \mid s)) \leq \delta. \tag{11}$$

The constraint in (11) requires that, on average over states visited by π, the new policy π' remains inside a Wasserstein ball of radius δ around the old policy. A related problem is, for $\lambda \geq 0$, to consider a Wasserstein penalty:

$$\max_{\pi'(\cdot|s)} \; \mathbb{E}_{a\sim\pi'(\cdot|s)}[A^{\pi}(s,a)] \; - \; \lambda\, \mathsf{W}(\pi'(\cdot \mid s),\, \pi(\cdot \mid s))\,.$$

In this formulation one explicitly balances improving $\mathbb{E}_{a\sim\pi'(\cdot|s)}[A^{\pi}(s,a)]$ with staying close to π.

In discrete action spaces these state-wise subproblems can be handled explicitly. Fix a state s and denote $p(\cdot) = \pi(\cdot \mid s)$ and $p'(\cdot) = \pi'(\cdot \mid s)$ as probability vectors over the finite action set. Song et al. (2023) consider the optimisation

$$\max_{p'(\cdot)} \; \sum_{a} p'(a)\, A^{\pi}(s,a) \quad \text{s.t.} \quad \mathsf{W}(p'(\cdot), p(\cdot)) \leq \delta,$$

with W chosen as either a 1-Wasserstein distance or an entropy-regularised approximation. They derive an explicit policy update conditional on the optimal dual variable, which itself is obtained from a one-dimensional optimisation problem and does not require specifying a particular parametric form for π'. The resulting update redistributes probability mass from low-advantage to high-advantage actions while paying a transport cost determined by $c(a, a')$. For a reader used to KL trust regions, the key difference is that the update is not only asking "how much probability should move?", but also "between which actions is that movement cheap enough to allow?"

Their Wasserstein Policy Optimisation (WPO) and Sinkhorn Policy Optimisation (SPO) algorithms implement these per-state updates across batches of states, directly at the level of discrete distributions, and establish monotonic-improvement guarantees under assumptions analogous to those used in TRPO. The 2023 paper further proves global convergence to an optimal policy (Theorem 5 of Song et al. 2023), strengthening the local guarantees of the 2022 version.

Compared to a KL ball, a Wasserstein ball makes the notion of a "small" policy change depend explicitly on the ground cost: moving probability mass between similar actions incurs a low cost, while moving mass between very different actions incurs a high cost. This allows domain knowledge about the action space (for example, which torques or control inputs are close in effect) to enter directly through c. For a practitioner, this is the main attraction of Wasserstein trust regions: the constraint can reflect the physical or semantic structure of the action space, rather than only probability ratios.

In continuous action spaces, the constraint in (11) becomes an infinite-dimensional problem over couplings between densities. Terpin et al. (2022) study trust regions defined by optimal transport discrepancies (which include Wasserstein distances) and propose OT-TRPO, an extension of TRPO to continuous actions. When policies are Gaussian and the ground cost is quadratic, W_2 has a closed-form expression in terms of means and covariances. Using convex duality, they show that the constrained optimisation problem can be reduced to a one-dimensional convex optimisation over the Lagrange multiplier associated with the OT constraint and derive an analytic expression for the optimal update of the Gaussian parameters given the dual solution. In practice this yields a TRPO-like update for the mean and covariance of the Gaussian policy that respects

the Wasserstein geometry of the action space, with computational cost comparable to standard second-order policy-gradient methods and without explicit computation of transport plans.

Beyond finite action spaces, Song et al. (2023) also discuss how to extend Wasserstein trust regions to continuous actions without introducing a fixed hand-crafted discretisation of the action space. Their approach first solves the WPO/SPO target distribution on a finite sampled support set drawn from the current policy, then trains an Implicit Quantile Network (IQN; Dabney et al. 2018a) to match this target via quantile regression. The IQN serves as a continuous generative model for the updated policy, allowing WPO/SPO-style updates to be applied in settings where the action space is large or continuous.

Wasserstein Policy Regularisation (WPR; Na et al. 2026) brings the same trust-region intuition to reinforcement learning from human feedback (RLHF), but with a different anchor point. Instead of keeping the policy close to the previous iterate, WPR keeps it semantically close to a reference policy under an entropy-regularised Wasserstein penalty. A dual-variable reformulation makes that penalty compatible with PPO-style clipped updates. A complementary DRO view of RLHF alignment appears in Section 7.4.

More recent work views Wasserstein trust regions through the lens of proximal updates and gradient flows. Pfau et al. (2025) derive a policy optimisation method by discretising a Wasserstein gradient flow of the expected-return functional (see Section 4.4). Their gradient-flow derivation yields a qualitatively different update direction from standard policy gradients: actions are weighted by the gradient of the action-value function rather than by a score-function estimator. In practice, however, the trust region in their implemented algorithm is enforced via a KL penalty rather than a Wasserstein constraint, so the theoretical Wasserstein perspective and the practical algorithm differ.

Across these variants, the underlying move is the same: replace the KL geometry of TRPO by a Wasserstein geometry, and measure the size of an update by the transport cost of moving one action distribution into another under the chosen ground cost.

4.3 Behaviour-guided regularisation in latent spaces

The per-state action view is not always the right one. In many applications the relevant object is a whole trajectory, or a learned summary of it, rather than an instantaneous action distribution. Given a behavioural embedding map $\Phi : \mathcal{T} \to \mathcal{E}$ that extracts such a summary, each policy π induces a *behavioural distribution* $\mu_\pi = \Phi_\# P_\pi$ on the latent space $\mathcal{E}$. The workflow is: choose the behavioural aspect that matters, encode trajectories into that space, and compare the resulting distributions there. A generic behaviour-regularised policy optimisation problem is then

$$\max_\theta \; \mathcal{L}(\theta) \; + \; \beta\, \mathcal{R}\big(\mu_{\pi_\theta}, \mu_{\mathrm{ref}}\big), \tag{12}$$

where $\mathcal{L}(\theta)$ is a standard policy objective, μ_{ref} is a reference behavioural distribution, $\mathcal{R}$ is a discrepancy measure, and the sign of β determines whether the regulariser attracts (imitation, safety) or repels (exploration, diversity). Non-OT choices for $\mathcal{R}$ include moment penalties, MMD, or kernel-determinant objectives such as DvD (Parker-Holder et al., 2020). These approaches typically rely on simple moment comparisons or densities and do not exploit the full geometry of the empirical distributions when supports differ or embeddings are not injective.

Pacchiano et al. (2020) propose to equip the family of behavioural distributions $\{\mu_\pi\}$ with a Wasserstein geometry and to use entropy-regularised Wasserstein distances W_ε (Section 2.2) as behavioural regularisers. Their starting point is to regard a policy as a distribution over behaviours, and to define the dissimilarity between two policies π and b via $\mathsf{W}_\varepsilon(\mu_\pi, \mu_b)$ under a ground cost c on $\mathcal{E}$. Because W_ε is differentiable and admits a dual formulation in terms of test functions on $\mathcal{E}$, it is amenable to stochastic-gradient methods.

In the Behaviour-Guided Policy Gradient (BGPG) algorithm, the policy objective is

$$F(\theta) \; = \; \mathcal{L}(\theta) \; + \; \beta\, \mathsf{W}_\varepsilon\big(\mu_{\pi_\theta}, \mu_b\big),$$

with $\mathcal{L}(\theta)$ taken to be a standard surrogate such as (10) and b a reference policy. The Wasserstein term acts as a behavioural trust region in the latent space: it discourages updates that move the distribution of

embeddings μ_{π_θ} too far from μ_b in a single step. Crucially, the metric is defined on the behaviour space $\mathcal{E}$ rather than on actions, so the regulariser respects the geometry of the chosen behavioural summary. Pacchiano et al. (2020) show that using smoothed Wasserstein distances is advantageous when the embedding map Φ is not injective, in which case likelihood-based divergences such as KL are difficult to apply. The conceptual shift is that two policies are now considered close when they *behave* similarly in the embedding space, even if their action probabilities differ sharply at individual time steps.

Algorithmically, BGPG uses the dual formulation of W_ε to learn *behavioural test functions* on $\mathcal{E}$, which assign scores to embeddings and can be evaluated on trajectories. These test functions play a similar role to discriminators in adversarial methods, but are derived from the OT dual and updated jointly with the policy using stochastic gradient descent in a reproducing-kernel Hilbert space. The scores are then added as shaping terms to the task reward, steering policy optimisation towards or away from specific regions of behaviour space.

A companion algorithm, Behaviour-Guided Evolution Strategies (BGES), applies the same principle to evolution strategies. In BGES, parameter perturbations are sampled from a Gaussian distribution, and each perturbation induces a behavioural distribution over embeddings. The fitness function is augmented with an entropic Wasserstein term that measures the distance between the perturbed policy's behavioural distribution and a reference distribution, so that perturbations producing behaviours closer to the desired reference receive higher fitness scores.

What matters in BGPG and BGES is that OT no longer acts on actions but on *behaviours*. Once the embedding map Φ and ground cost c on $\mathcal{E}$ are fixed, the Wasserstein term quantifies how much the policy's behavioural distribution must move to match a reference. This produces a family of regularisers that naturally handles non-injective embeddings and partial behavioural summaries, where action-level KL penalties are often a poor fit.

4.4 Wasserstein gradient flows and natural gradients

The trust-region and behaviour-guided methods above still start from a standard policy objective and then add OT as a penalty or constraint. A more ambitious viewpoint is to take OT as the underlying *geometry* of policy space itself and derive update rules from that geometry.

To make this precise, it is helpful to think of a policy not only as a parameter vector, but as a probability measure. Depending on the context, this measure can be taken over actions (conditioned on state) or over whole trajectories. Let $\mathcal{X}$ denote the space on which we represent policies as measures (for example, a trajectory space $\mathcal{T}$ or a state–action space $\mathcal{S} \times \mathcal{A}$), and let $\mathcal{P}(\mathcal{X})$ be the space of probability measures on $\mathcal{X}$. A functional

$$F : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$$

assigns a cost to each such measure; for policy optimisation, F could be the negative expected return under the corresponding policy. A Wasserstein gradient flow of F is then a curve $(\mu_t)_{t\geq 0}$ in $\mathcal{P}(\mathcal{X})$ that evolves according to the steepest descent of F with respect to the 2-Wasserstein metric. A convenient discrete-time characterisation of such flows is given by the Jordan–Kinderlehrer–Otto (JKO) scheme (Ambrosio et al., 2008):

$$\mu_{k+1} = \operatorname*{argmin}_{\mu \in \mathcal{P}(\mathcal{X})} \Big\{ F(\mu) + \frac{1}{2h}\, \mathsf{W}_2^2(\mu, \mu_k) \Big\}, \tag{13}$$

where $h > 0$ is a step size and W_2 is the 2-Wasserstein distance with respect to a ground metric on $\mathcal{X}$. Each step chooses a new measure that decreases F while staying close, in W_2, to the previous measure μ_k. This is the measure-valued analogue of an implicit gradient step, but with distance measured in Wasserstein space rather than in Euclidean space. One can read (13) as a proximal policy update in which "do better" means reduce the objective F, while "do not move too far" is measured by transport cost between the old and new policy-induced measures.

Zhang et al. (2018) adapt this construction to RL by applying Wasserstein gradient flows on suitable spaces of policy-related probability measures. In their formulation, μ_k is the policy parameter distribution or

energy-based policy distribution (e.g. per-state action distribution), and the minimisation in (13) is achieved by particles with the policy approximated by a parameterised state-conditioned neural network. The energy functional F is defined based on the expected return modelled by a Q-network, and the Wasserstein term acts as a regulariser that discourages abrupt changes in the policy update. Conceptually, the picture is closer to moving a cloud of policy particles smoothly toward higher return than to taking an unconstrained gradient step in parameter space. Earlier, Richemond & Maginnis (2017) linked Wasserstein gradient-flow perspectives to the Fokker–Planck equation and Langevin-type dynamics in policy space, providing a continuous-time interpretation of policy updates as diffusion processes in Wasserstein geometry; however, their analysis is restricted to the bandit (single-step) setting under strong regularity assumptions.

For parametric policies π_θ, the influence of Wasserstein geometry can be expressed in terms of a matrix that plays the role of a metric tensor on parameter space. Drawing on the general theory of Wasserstein information geometry (Li & Montúfar, 2018), one can define a Riemannian metric on the parameter manifold by pulling back the W_2 metric defined on the space of measures. In this framework, Moskovitz et al. (2021) introduce a *Wasserstein information matrix* $G(\theta)$ for policies and define the Wasserstein natural gradient as

$$g^W = G(\theta)^{-1}\,\nabla_\theta \mathcal{L}(\theta),$$

where $\mathcal{L}(\theta)$ is a policy objective (for example, the surrogate in (10)) and $\nabla_\theta \mathcal{L}(\theta)$ is its Euclidean gradient. Informally, $G(\theta)$ captures how small changes in parameters θ affect the W_2 distance between nearby policy-induced measures; the direction g^W is the steepest-ascent direction when distances between policies are measured in Wasserstein space rather than in parameter space. This is the Wasserstein analogue of what the Fisher information matrix does for classical natural gradients: it rescales parameter directions according to how strongly they deform the policy distribution.

Computing $G(\theta)$ exactly is infeasible except in simple cases. Moskovitz et al. (2021) therefore propose low-rank and kernel-based approximations, leading to practical algorithms for Wasserstein natural gradients in both policy-gradient (WNPG) and evolution-strategy (WNES) settings. They report improvements in computational cost and performance, and highlight an advantage over Fisher natural gradients in settings where the optimal solution is deterministic. This connects back to the trust-region view: including a Wasserstein penalty in the objective, as in (11), can be seen as inducing a local metric, and Wasserstein natural gradients provide an explicit approximation to the corresponding Riemannian geometry.

The proximal formulation (13) also clarifies how Wasserstein gradient flows relate to the OT-based trust-region methods of Section 4.2. A JKO step is a proximal update in Wasserstein space. Methods such as WPO/SPO (Song et al., 2023) and Wasserstein policy optimisation (Pfau et al., 2025) can therefore be read as discrete-time approximations to policy optimisation in that geometry. For WPO/SPO the link is direct, because the algorithms use Wasserstein or Sinkhorn trust regions explicitly. For Pfau et al. (2025), the link is looser: the derivation is Wasserstein, but the implemented algorithm enforces its trust region with KL regularisation. The broader lesson is that Wasserstein trust regions, behavioural regularisers, and Wasserstein natural gradients are not separate ideas so much as different ways of choosing a geometry on policy space and then asking updates to stay small in that geometry while still improving return. The same gradient-flow viewpoint reappears in the offline setting in Section 6.

4.5 Wasserstein-trained critics

So far OT has acted on the policy side of actor–critic learning: it constrained the actor directly or supplied the geometry in which policy updates were defined. It also appears on the critic side. Since actor–critic methods plug critic-based advantage estimates into the surrogate objective (10), any Wasserstein loss used to train the critic indirectly changes the policy update as well.

Wasserstein distances appear in critic learning in two distinct ways. The first concerns returns in distributional RL, where the full return distribution $Z^\pi(s,a)$ is modelled rather than just its mean. The second concerns the value function, where critics model posterior distributions over Q-values updated via Wasserstein barycentres.

Distributional RL in Wasserstein and Sinkhorn metrics. Instead of estimating the mean return for each state-action pair, distributional RL models the distribution of returns to capture the inherent randomness.

Because return distributions are one-dimensional, Wasserstein distances can be computed through quantile comparisons, making OT a computationally efficient and geometrically meaningful loss for distributional critics. Distributional Bellman operators are contractive in suitable Wasserstein metrics under standard assumptions, motivating losses that measure the Wasserstein discrepancy between predicted and target return distributions. With the action-value distribution Z^π, the distributional Bellman operator is

$$\mathcal{T}^\pi Z^\pi(s,a) \stackrel{D}{=} R(s,a) + \gamma Z^\pi(s',a'), \quad s' \sim P(\cdot \mid s,a),\ a' \sim \pi(\cdot \mid s'),$$

and the distance between current and target Bellman return distributions is measured by

$$d_{\sup}(Z_{\theta_z}, \mathcal{T}^\pi Z_{\bar{\theta}_z}) = \sup_{s,a} \mathsf{W}_p(Z_{\theta_z}(s,a), (\mathcal{T}^\pi Z_{\bar{\theta}_z})(s,a)),$$

where $Z_{\bar{\theta}_z}$ is the target for stability in practice. The practical point is that the critic is no longer asked only for the mean return, but for the whole distribution of plausible returns at each (s,a). In practice, QR-DQN (Dabney et al., 2018b) approximates the W_1 loss via quantile regression on one-dimensional return distributions, providing a distributional alternative to scalar (mean-return) critics. More recently, SinkhornDRL (Sun et al., 2024) replaces the 1D quantile-based surrogate with a Sinkhorn divergence between particle-based distributional estimates, enabling multi-dimensional reward settings and improved numerical stability through entropic regularisation; debiasing (subtracting self-transport terms) preserves metric properties and meaningful geometry. Both methods provide theoretical contraction guarantees and empirical improvements over non-distributional baselines. The relevant Bellman or projected Bellman operators have method-specific contraction or non-expansion guarantees, depending on the metric, projection, and representation. Distributional critics also enable OT-based off-policy evaluation, using Wasserstein distances between predicted and target return distributions (Qi et al., 2025).

Multi-dimensional return distributions are particularly relevant for safe RL with constraints, multi-objective RL, and exploration via learned joint reward distributions.

Wasserstein barycentres for Q-posteriors. Another way of using Wasserstein distances is to propagate uncertainty of Q-value estimates through Bellman updates instead of modeling return randomness. Metelli et al. (2019) propose Wasserstein Q-learning (WQL), in which Q-values are represented as probability distributions (posteriors) rather than point estimates. For each state–action pair (s,a), the critic maintains a distribution $Q(s,a)$ and updates it via a 2-Wasserstein barycentre between the previous estimate and a target distribution. More precisely, at time t the new Q-posterior at (s_t,a_t) is defined as the 2-Wasserstein barycentre

$$Q_{t+1}(s_t,a_t) = \operatorname*{argmin}_{Q} \Big\{ (1-\alpha_t)\, \mathsf{W}_2^2\big(Q, Q_t(s_t,a_t)\big) + \alpha_t\, \mathsf{W}_2^2\big(Q, r_{t+1} + \gamma V_t(s_{t+1})\big) \Big\},$$

where α_t is a learning rate, $V_t(s)$ is a value-posterior obtained as a 2-Wasserstein barycentre of $\{Q_t(s,a) : a \in \mathcal{A}\}$, and $r_{t+1} + \gamma V_t(s_{t+1})$ denotes the distribution of the one-step reward plus the discounted next-state value. In Gaussian and particle-based approximations this minimiser has a closed form, so that uncertainty is propagated across the Bellman update by 2-Wasserstein barycentres with modest computational overhead. This update can be read as a distributional analogue of TD learning: instead of averaging two scalar targets, it takes a Wasserstein barycentre between the old Q-posterior and the new Bellman target distribution.

Likmeta et al. (2023) extend this idea to continuous-action actor–critic methods in Wasserstein Actor–Critic (WAC). They maintain Gaussian Q-posteriors and update them using W_2 to target distributions, similarly to WQL. The actor is then updated according to an optimistic estimate of Q, derived from the upper tail of the Q-posteriors, so that Wasserstein-based uncertainty quantification guides exploration in action space. Recent work extends this to MMD barycentres as an alternative to Wasserstein for model-free uncertainty propagation (Roy et al., 2025).

Even without explicit distributional modelling, Wasserstein regularisation can improve the stability of standard critics. The WAVE algorithm (Baheri et al., 2025) augments the usual TD loss with a small entropy-regularised Wasserstein term between empirical Q-value distributions at successive iterations:

$$L(\theta) = L_{\mathrm{TD}}(\theta) + \lambda_k\, \mathsf{W}_\varepsilon\big(\hat{Q}_k, \hat{Q}_{k-1}\big),$$

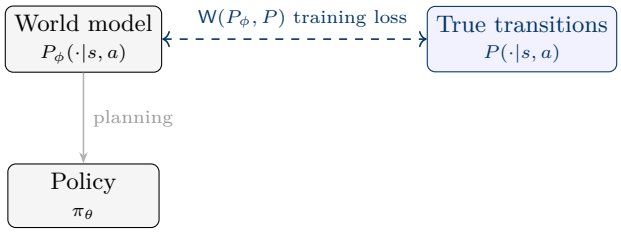

Figure 9: OT in model-based RL (Section 5). A learned world model $P_\phi(\cdot \mid s, a)$ is trained to match the true transitions under a Wasserstein loss that upper-bounds worst-case value error. Bisimulation metrics embed this comparison into a fixed-point state metric (Section 5.3).

where L_{TD} is a standard TD error, $\hat{Q}_k$ and $\hat{Q}_{k-1}$ are empirical distributions of Q-values under the critic at iterations k and $k-1$, and λ_k is an adaptive regularisation weight based on recent performance. The transport term discourages abrupt changes in Q-values from one update to the next and leads to smoother critic and policy dynamics.

For policy optimisation, these methods modify not the policy update rule itself, but the critic that supplies the value estimates. Wasserstein distances govern how value distributions evolve with training, and hence indirectly regularise the gradients used for policy updates.

4.6 Summary

Across the methods in this section, OT acts on three kinds of policy-induced objects: action distributions, behavioural embeddings, and value or return distributions. The details differ, but the underlying move is the same. KL-style methods ask whether the new policy assigns similar probability to the same actions; OT asks how much probability mass must actually move, and through what geometry, to turn one distribution into another.

That shift leads to four main uses: trust regions on action distributions, regularisation of trajectory embeddings, gradient-flow or natural-gradient updates in Wasserstein geometry, and critic training based on return or Q-value distributions. OT in policy optimisation is therefore not a single algorithm but a family of ways to redefine what counts as a small, structured update. When the action geometry is weak or uninformative, KL/PPO-style updates are usually the simpler baseline.

5 Model-based RL and Wasserstein world models

Model-based RL differs from model-free RL in that it relies on a learned, or nominal, transition model $P(\cdot \mid s, a)$ for planning, value expansion, or policy optimisation. This can greatly improve sample efficiency, but it introduces a new failure mode: if the learned world model is wrong, multi-step rollouts and Bellman backups can amplify small local errors into large control errors. OT is useful here because it measures model error in terms of how next-state distributions or trajectory distributions must move, under a chosen state-space geometry, rather than only through pointwise likelihood or regression losses (Figure 9). The guiding question in this section is therefore not just "is the model accurate?", but "is it accurate in the ways that matter for control?"

This section focuses on OT as a tool for *world modelling*. We first look at value-aware losses between true and learned next-state distributions $P(\cdot \mid s, a)$ and $P_\phi(\cdot \mid s, a)$. We then turn to latent-model objectives, where OT is used to learn compressed MDPs that preserve control-relevant structure, often with bisimulation-style guarantees. Robust control via Wasserstein ambiguity sets is closely related, but conceptually different: there OT hedges against transition uncertainty rather than fitting a model to data, and we discuss it in Section 7.

5.1 Model-based RL in brief

A standard model-based pipeline learns $P_\phi(\cdot \mid s, a)$ from data and then uses the model for planning or policy optimisation, for example through model predictive control, Monte Carlo tree search, Dyna-style imagined rollouts, or value expansion. The basic challenge is *compounding error*: one-step prediction may look accurate on average, yet the induced multi-step rollouts may still distort value estimates enough that the policy learns to exploit model flaws. A useful intuition is that small local errors can become strategically important when planning repeatedly feeds the model's own predictions back into itself. OT-based objectives address this by replacing purely pointwise prediction losses with discrepancies that reflect both state-space geometry and the control relevance of prediction errors.

5.2 Value-aware model learning and Wasserstein losses

In model-based RL, one learns an approximate transition model $P_\phi(\cdot \mid s, a)$ (and possibly a reward model R_ϕ) from data and uses it for planning or to generate fictitious rollouts. A standard approach is maximum likelihood estimation (MLE), which minimises the KL divergence from the empirical transition distribution to the model:

$$\hat{P}(\cdot \mid s, a) \;=\; \operatorname*{argmin}_{P \in \mathcal{M}} \; \mathrm{KL}\big(P_n(\cdot \mid s, a) \,\|\, P(\cdot \mid s, a)\big) \;=\; \operatorname*{argmin}_{P \in \mathcal{M}} -\frac{1}{n} \sum_{i=1}^{n} \log P(s_i' \mid s, a),$$

where $P_n(\cdot \mid s, a) = \frac{1}{n}\sum_{i=1}^{n} \delta_{s_i'}$ is the empirical measure of observed next states. This objective treats all prediction errors equally, regardless of how important they are for the control problem. Predicting a visually different but value-equivalent next state is penalised just as much as predicting a state that changes the best action, even though those two mistakes matter very differently for control.

Farahmand et al. (2017) argued that model-learning objectives should instead be *value-aware*: not all aspects of the environment are helpful to find an optimal policy, and the decision problem is reflected by the reward function. Errors that significantly change the value function should be penalised more than errors that are irrelevant for control. In plain terms, the model should care most about getting the consequences of decisions right. One way to formalise this is to measure the worst-case discrepancy in value across a class of test functions. For a given state–action pair (s, a), define

$$\mathcal{E}(s, a) \;=\; \sup_{V \in \mathcal{V}} \left| \mathbb{E}_{s' \sim P(\cdot \mid s, a)}[V(s')] - \mathbb{E}_{\hat{s}' \sim P_\phi(\cdot \mid s, a)}[V(\hat{s}')] \right|^2,$$

where $\mathcal{V}$ is a class of value functions. The value-aware loss then aggregates $\mathcal{E}(s, a)$ over (s, a) drawn from the data distribution; intuitively, this loss is small if the learned model and the true dynamics produce similar expectations for all test functions in $\mathcal{V}$. The role of the supremum is important: it asks for the *worst* value function in the class, so the model is penalised according to the largest control-relevant mistake it could induce at (s, a).

A natural choice for $\mathcal{V}$ is the class of Lipschitz functions with respect to a chosen state metric d. By the Kantorovich–Rubinstein duality (Section 2.2), the supremum over 1-Lipschitz functions recovers the W_1 distance, so that

$$\mathcal{E}(s, a) \;=\; \mathsf{W}_1^2\big(P(\cdot \mid s, a),\; P_\phi(\cdot \mid s, a)\big).$$

Asadi et al. (2018) formalised this connection: when $\mathcal{V}$ consists of C-Lipschitz functions (where C depends on the MDP discount and reward bound), minimising the value-aware objective $\mathcal{E}(s, a)$ is equivalent to minimising the squared W_1 distance between the true and learned next-state distributions. This provides a direct link between model-based RL and OT: training P_ϕ to minimise a Wasserstein distance with a value-relevant state metric is equivalent to training it to minimise the worst-case squared value error over a Lipschitz class of value functions. The key intuition is that the state metric weights model errors by their impact on downstream value. This is the point at which OT becomes more than a generic discrepancy: once the metric reflects control similarity, Wasserstein model fitting becomes a way of fitting the model for planning rather than merely for prediction.

5.3 Bisimulation metrics and Markov-structured transport

Bisimulation metrics formalise when two states are "similar for control": they compare both immediate rewards and the similarity of next-state distributions under each action. A classic definition is the (unique) fixed-point of a contraction mapping (Ferns et al., 2004). For a candidate state metric d and a chosen ground metric on $\mathcal{S}$ (often d itself), one form is

$$d(s, s') = \max_{a \in \mathcal{A}} \Big((1-\gamma) \left| R(s,a) - R(s',a) \right| + \gamma \, \mathsf{W}_1\big(P(\cdot \mid s,a),\, P(\cdot \mid s',a)\big) \Big),$$

with entropy-regularised variants used in practice for scalability (Kemertas & Jepson, 2022). Such metrics upper-bound the difference in optimal values between s and s' and motivate representation learning objectives that collapse bisimilar states. One should read this fixed-point equation as balancing two questions at once: do the states give similar immediate reward, and do they lead to similar future-state distributions under every action?

Recent work connects these ideas to OT between *stochastic processes* and shows how to incorporate Markov structure directly into the coupling class (Moulos, 2021; Calo et al., 2024). Instead of arbitrary couplings over full trajectories, one restricts to *Markovian couplings* between two Markov chains with transition kernels $\mu(\cdot \mid x)$ and $\nu(\cdot \mid y)$. Concretely, a Markovian coupling is a kernel $\kappa(\cdot \mid x, y)$ on $\mathcal{X} \times \mathcal{Y}$ such that its marginals agree with the original transitions:

$$\sum_{y'} \kappa(x', y' \mid x, y) = \mu(x' \mid x), \qquad \sum_{x'} \kappa(x', y' \mid x, y) = \nu(y' \mid y).$$

Unlike standard couplings, Markovian couplings ensure that the pairing at time $t+1$ depends only on the paired states at time t, respecting the sequential structure of MDPs. This is exactly the extra structure missing from ordinary trajectory couplings: the transport itself has to respect the Markov evolution of the process. Writing $\Pi_{x,y}$ for the set of such admissible couplings at the pair (x, y), discounted OT costs can be characterised by Bellman-style equations of the form

$$V^\star(x, y) = c(x, y) + \gamma \inf_{\kappa \in \Pi_{x,y}} \sum_{x', y'} \kappa(x', y' \mid x, y)\, V^\star(x', y').$$

When the one-step cost c is chosen to reflect reward mismatch, these dynamic-programming characterisations align closely with bisimulation fixed points and suggest practical solvers based on value iteration, including entropy-regularised variants with Sinkhorn-style updates (Calo et al., 2024). The Bellman form is not accidental: it says that transport between processes should itself respect the recursive structure of control problems, not just match isolated snapshots of state.

Algorithmically, this viewpoint provides OT computations that respect temporal/Markov structure and clarifies why OT-like objectives in latent world models can support bisimulation-style guarantees: the transport discrepancy is computed over transitions in a way that is consistent with the dynamics, rather than over unordered state sets.

5.4 Wasserstein auto-encoded and OT-based latent world models

The bisimulation perspective motivates a more ambitious goal: learning an entire latent MDP whose dynamics are trained via explicit OT objectives, a departure from reconstruction/likelihood-based latent world models such as (Corneil et al., 2018; François-Lavet et al., 2019; Hafner et al., 2021). WAE-MDPs (Delgrange et al., 2023a) combine the framework of Wasserstein autoencoders with MDPs. They introduce an encoder ϕ that maps high-dimensional states to latent states, a decoder ψ that reconstructs states from latent states, and a latent transition model $\bar{P}$. The key idea is to train this latent MDP by minimising a penalised optimal-transport distance between *trace* distributions: the distribution of trajectories generated by the original policy in the real environment, and the distribution of trajectories generated by the distilled policy in the latent MDP and then decoded back to the original space. The intuition is that the latent model should not merely reconstruct observations; it should reproduce the behaviourally relevant trajectories of the original system.

Concretely, WAE-MDPs augment the VAE-MDP objective with OT-based terms: a global Wasserstein penalty on trace (trajectory) distributions $L_{\xi_{\bar{\pi}}}$, together with local losses that encourage the latent reward L_R and transition functions to approximate the original dynamics L_P. As part of the overall loss function, the local transition loss is defined as

$$L_P(\phi, \bar{P}) = \mathbb{E}_{(s,a)\sim\xi_{\bar{\pi}}} \mathsf{W}_1\big(\phi_\# P(\cdot \mid s, a),\ \bar{P}(\cdot \mid \phi(s), a)\big),$$

where $\xi_{\bar{\pi}}$ is the stationary distribution of the policy $\bar{\pi}(\cdot|\phi(s))$. This term compares the real next-state distribution after encoding with the latent model's predicted next-state distribution. If it is small, then stepping once in latent space means roughly the same thing as stepping once in the real environment and then encoding. The resulting latent model is shown to satisfy bisimulation-like properties: states that are functionally equivalent for control are mapped close together in latent space. This, in turn, enables formal verification and safety analysis to be performed in the latent MDP, with guarantees that transfer back to the original environment. The high-level picture is that the latent state is not just a compressed observation; it is meant to be a compressed control state. A closely related extension, WAE-PCN (Delgrange et al., 2023b), combines Wasserstein autoencoded latent models with Pareto-conditioned networks for multi-objective RL.

Optimal Transport MDPs (OT-MDPs) (Röpke et al., 2025) generalise and simplify this idea. They define a generic regularised OT objective between trajectories in the original MDP and trajectories in a learned latent MDP to learn the encoder, latent transition model and reward model:

$$\inf_{\phi,\bar{P},\bar{R}} \mathbb{E}_{\tau\sim\xi_{\bar{\pi}}}[c(\tau, \bar{\tau})] + \lambda(L_{\xi_{\bar{\pi}}} + L_P + L_R)$$

where c can be chosen with more freedom than the fixed reconstruction cost $c_{\mathrm{WAE}}(\tau, \bar{\tau}) = d_S(s, \psi(\bar{s})) + d_S(s', \psi(\bar{s}'))$ used by WAE-MDPs. OT-MDPs define a general cost of the form

$$c(\tau, \bar{\tau}) = h\big(f(s), g(\bar{s})\big),$$

where f and g are feature maps on real and latent states and h is a distance or divergence in feature space. In a specific instantiation, $c_{\mathrm{value}}(\tau, \bar{\tau}) = |V(s) - \bar{V}(\bar{s})|$ uses value functions directly, yielding a cost that prioritises control-relevant similarity. The extra flexibility matters because it lets the modeller decide what two trajectories should count as "the same" for control purposes: visual similarity, feature similarity, or similarity in downstream value. The OT-MDP framework then gives conditions on c under which the learned latent model is bisimilar to the original MDP, and shows how planning tools such as model-based value expansion and Monte Carlo tree search can be run in the latent space. The OT objective is not just a training loss here; it is the quantity that controls how faithfully values and planning decisions transfer from the latent model back to the original environment. In practice, WAE-MDPs enforce 1-Lipschitzness with a gradient penalty, whereas OT-MDPs discretise the relaxed latent state and enforce 1/2-boundness through an additional penalty.

For control, WAE-MDPs and OT-MDPs use optimal transport to train world models that preserve the structure of the original MDP in a way that is relevant for value and safety, rather than merely matching raw observations. Policies are then optimised in these latent models using standard or OT-based policy-optimisation methods. The practical takeaway is that OT is being used as a filter on what kind of model fidelity matters: not pixel-level fidelity, but control-relevant fidelity. The closer the latent model is to the true environment in that Wasserstein sense, the more reliable the resulting policy improvements are.

5.5 Summary

OT enters model-based RL in three main ways: as a value-aware discrepancy between learned and true next-state laws, as the metric inside bisimulation-style state comparisons, and as an explicit trajectory-matching objective in latent world models.

Across these methods, the central modelling choice is the *metric structure* used to compare next-state distributions or trajectory traces. In value-aware model learning, this choice appears as the state metric that defines 1-Lipschitz test functions (hence the implied W_1 geometry). In latent-model objectives, it appears as

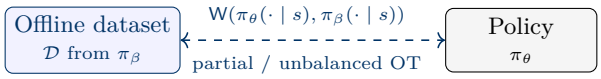

Figure 10: OT in offline RL (Section 6). The agent learns entirely from a fixed dataset $\mathcal{D}$ collected from the behaviour policy π_β. OT enters as a behaviour constraint $\mathsf{W}(\pi_\theta(\cdot \mid s), \pi_\beta(\cdot \mid s))$ that keeps the learned policy geometrically close to the data-generating policy with both induced distributions approximated by collected trajectories (Section 6.2). Partial and unbalanced OT relax mass conservation to handle suboptimal demonstrations.

the trajectory feature maps f, g and the cost c used in the OT alignment. In high-dimensional observation spaces, these costs are typically defined in learned feature spaces rather than raw pixel space.

Ambiguity-set methods in Section 7 use the same geometry for a different purpose: hedging against shifted transition laws rather than fitting a model to data. The same modelling choices also reappear in offline settings, where learned dynamics are especially exposed to distributional shift and compounding error (Section 6).

Several extensions are suggested by the OT tools reviewed in this survey. First, entropic OT/Sinkhorn discrepancies provide scalable minibatch objectives for matching conditional next-state laws $P(\cdot \mid s, a)$ via samples, and can be combined with learned feature costs to handle high-dimensional observations. Second, partial and unbalanced OT are well suited when transition data exhibit missing outcomes, outliers, or systematic mass mismatch (for example, logged data that omit rare but important transitions), and may yield more robust world models than strict balanced matching.

A newer direction makes the OT–flow connection operational for *shifted dynamics*. Composite Flow Matching for RL with Shifted Dynamics Data (CompFlow) uses flow-matching constructions to model transition dynamics under distribution shift and estimates a *dynamics gap* through Wasserstein distances between learned flow models. That gap is then used to guide data collection and exploration toward regions where source-domain data are least informative for the target dynamics (Kong et al., 2025).

6 Offline policy optimisation and regularisation

Offline RL restricts the agent to a fixed dataset of transitions, with no environment interaction during learning (Figure 10). The central difficulty is distributional shift: once the learned policy chooses actions that are poorly represented in the data, the critic must extrapolate into unsupported regions and its value estimates can become badly biased. Standard remedies keep the learned policy close to the data-generating behaviour π_β, but density-ratio penalties such as $\mathrm{KL}(\pi \| \pi_\beta)$ are a poor fit when supports differ, which could be too conservative and prevent further improvement beyond the behaviour policy, and lack of geometric awareness. An OT cost $\mathsf{W}(\pi(\cdot \mid s), \pi_\beta(\cdot \mid s))$, by contrast, remains informative as long as a ground cost $c(a, a')$ is defined on $\mathcal{A} \times \mathcal{A}$ and encourages generalisation. It therefore gives a geometric notion of staying close to the data: shifting mass between similar actions is cheap, while shifting it to qualitatively different actions is expensive. The intuition is the same as in policy trust regions, but the reason is different: here the goal is not mainly to stabilise optimisation, but to bias policy updates towards actions that remain geometrically close to what in the dataset.

6.1 Standard offline RL framework

In offline RL (Lange et al., 2012; Levine et al., 2020), the agent learns from a fixed dataset $\mathcal{D} = \{(s_t, a_t, r_t, s_{t+1})\}$ generated by an unknown behaviour policy π_β and cannot interact with the environment. Unlike imitation learning (Section 3), a reward function is available and the objective is to *improve* upon π_β, not merely replicate it. As π deviates from π_β, the agent increasingly queries Q-values in regions of $\mathcal{S} \times \mathcal{A}$ that are poorly covered by $\mathcal{D}$, leading to overestimation, compounding bootstrapping errors, and eventual policy collapse. A good informal picture is that the critic only knows the terrain where data have been collected; once the actor leaves that terrain, its value estimates become guesses.

Two broad ideas dominate offline RL. The first is to *constrain the policy*: keep the learned policy close enough to π_β that the actor does not query the critic far outside the support of the data (Fujimoto et al., 2019;

Kumar et al., 2019; Fujimoto & Gu, 2021; Siegel et al., 2020; Zhou et al., 2021; Brandfonbrener et al., 2021; Hepburn et al., 2025). Because π_β is unknown, it must first be estimated, usually by supervised learning. A common formulation is

$$\pi^* = \operatorname*{argmax}_{\pi} \mathbb{E}_{s\sim\mathcal{D}} \left[Q(s, \pi(s)) - \beta\, D\left(\pi(s), \pi_\beta(s)\right)\right], \tag{14}$$

where D denotes a divergence or regulariser, for example KL or MMD, and β controls the strength of the constraint. OT provides a natural alternative for D: measuring deviation via $\mathsf{W}(\pi(\cdot \mid s), \pi_\beta(\cdot \mid s))$ makes "small" deviations depend on the geometry of $\mathcal{A}$ and remains meaningful under support mismatch. In that sense, offline OT regularisation is the data-constrained counterpart of the Wasserstein trust-region idea from Section 4.2. The difference is that the reference policy is now the dataset's behaviour policy rather than the previous policy iterate.

The second idea is *pessimistic value estimation*: rather than constraining the actor directly, modify the Q-function so that out-of-distribution actions receive conservative values (Kumar et al., 2020; An et al., 2021; Kostrikov et al., 2022). The intuition is to make the critic suspicious of actions it has little evidence about. A general critic objective is

$$\mathcal{L}(\theta_q) = \mathrm{TDLoss}(Q_{\theta_q}) + \lambda\, \mathcal{P}\left(Q_{\theta_q}, \mathcal{D}\right), \tag{15}$$

where $\mathcal{P}$ enforces pessimism outside the support of $\mathcal{D}$. OT-based policy constraints already induce a form of implicit pessimism, because they stop the actor from moving too aggressively into unsupported regions; OT-based ambiguity sets over transitions (Section 7) can then make the critic targets themselves robust. Model-free and model-based versions both exist (Yu et al., 2020; Kidambi et al., 2020; Yu et al., 2021; Janner et al., 2022; Rigter et al., 2022; Hepburn & Montana, 2024; Wang et al., 2021; Zhang et al., 2021; Diehl et al., 2021), as do methods that combine explicit constraint with pessimistic estimation (Wu et al., 2019; Kostrikov et al., 2021; Dadashi et al., 2021b; Beeson & Montana, 2024). The rest of this section focuses on OT-based behaviour constraints and on partial or unbalanced formulations that relax exact matching.

6.2 OT-based policy constraints, partiality, and unbalanced formulations

The most common offline use of OT is as a *behaviour regulariser*: the learned policy is discouraged from selecting actions that are far, in an action-geometry sense, from those supported by the dataset. OTR-style reward labelling (Section 3.4) provides a complementary offline application, assigning pseudo-rewards via couplings rather than constraining the actor. In this section, the recurring question is therefore: should OT be used to keep the actor near the data, or to decide which parts of the data should matter most?

BRAC: a unifying framework. Wu et al. (2019) introduce *behaviour-regularised actor critic* (BRAC), a generic actor–critic template augments either the actor objective, the critic target, or both, with a divergence $D\big(\pi(\cdot \mid s), \pi_\beta(\cdot \mid s)\big)$. Writing Q_{θ_q} for the critic and π_θ for the actor, BRAC supports two modes: a *value penalty* mode, which subtracts $\alpha\, D(\pi_\theta(\cdot \mid s'), \pi_\beta(\cdot \mid s'))$ inside the Bellman target,

$$\min_{\theta_q}\ \mathbb{E}_{(s,a,r,s')\sim\mathcal{D},\, a'\sim\pi_\theta(\cdot|s')}\Big[\big(r + \gamma\big(Q_{\bar{\theta}_q}(s',a') - \alpha\, D(\pi_\theta(\cdot \mid s'), \pi_\beta(\cdot \mid s'))\big) - Q_{\theta_q}(s,a)\big)^2\Big],$$

where $Q_{\bar{\theta}_q}$ is the target critic updated slowly to improve stability of learning, and a *policy regularisation* mode, which adds the penalty only to the actor update,

$$\max_{\theta}\ \mathbb{E}_{s\sim\mathcal{D}}\Big[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\big[Q_{\theta_q}(s,a)\big]\ -\ \alpha\, D\big(\pi_\theta(\cdot \mid s), \pi_\beta(\cdot \mid s)\big)\Big]. \tag{16}$$

The divergence D is treated as a modular design choice; BRAC evaluates KL, kernel MMD, and Wasserstein distance. Replacing KL-type penalties by OT discrepancies (e.g. W_1, W_2, or Sinkhorn-based surrogates) makes the constraint *geometric*: moving probability mass between similar actions is cheap, while moving mass between dissimilar actions is expensive. This is exactly the kind of notion of closeness one wants when the action space has structure that likelihood ratios alone cannot express. The Wasserstein variant (BRAC-W) instantiates the Kantorovich–Rubinstein dual (7) on per-state action distributions, $\mathsf{W}(\pi(\cdot \mid s), \pi_\beta(\cdot \mid s)) =$

$\sup_{\|g\|_{\mathrm{Lip}}\le 1} \mathbb{E}_{a\sim\pi}[g(a)] - \mathbb{E}_{a\sim\pi_\beta}[g(a)]$, with a gradient penalty enforcing the Lipschitz constraint (Gulrajani et al., 2017). A central empirical finding is that simple behaviour regularisation is the essential ingredient for stable offline learning; many design complexities in prior work (adaptive penalty weights, large Q-ensembles) prove unnecessary. For the reader, BRAC is a useful anchor point: it is the clean baseline "stay near the data" framework, and the later OT methods mostly differ in how they define "near" or how much of the data they insist on matching.

The next three strands all depart from BRAC-style behaviour regularisation in different ways. Q-DOT replaces the adversarial Wasserstein critic by an explicit W_2 transport map. Gradient-flow approaches interpret the update itself through Wasserstein geometry rather than as a penalty around a baseline policy. Partial and unbalanced OT methods relax exact mass matching so that suboptimal or contaminated datasets need not be copied wholesale.

W_2-map regularisation (discriminator-free). When W_2 geometry is desired, Brenier's theorem (Section 2.2) implies that (under absolute continuity of the source and with the squared Euclidean cost) the optimal transport map is $T = \nabla\psi$ for a convex potential ψ, so that $\mathsf{W}_2^2(\mu,\nu) = \int \|x - \nabla\psi(x)\|_2^2 \,\mathrm{d}\mu(x)$. Previous offline-RL methods that use Wasserstein regularisation typically estimate the discrepancy through a dual critic or another adversarial component. Q-DOT (Omura et al., 2025) instead aims for W_2 geometry without a discriminator by parameterising a convex potential with input-convex neural networks (ICNNs). The learned policy is represented as the pushforward $\pi = \nabla\psi_{\#}\pi_\beta$ of the behaviour policy through the gradient of that potential, and the regularised actor objective becomes

$$\mathbb{E}_{(s,a)\sim\mathcal{D}}[Q_\phi(s, \nabla_a\psi(s,a)) - \alpha\|a - \nabla_a\psi(s,a)\|_2^2].$$

This gives a discriminator-free alternative to adversarial Wasserstein penalties. In addition, the method trains a separate Advantage-Weighted Regression (AWR) policy to propose actions that need not appear explicitly in the offline dataset. Both BRAC-W and Q-DOT evaluate OT objectives over minibatches rather than full distributions. The empirical Wasserstein distance is biased in finite samples (Bellemare et al., 2017); Fatras et al. (2020) show that taking expectations of per-minibatch Sinkhorn losses yields an unbiased gradient estimator, providing a practical remedy when scaling these methods to large datasets.

Value gradient flow without regularisation. Consider the MaxEnt RL objective $\mathbb{E}_\pi[Q(s,a) - \alpha\log(\pi(a \mid s))]$, whose unique maximiser is the Boltzmann policy $\pi_Q^*(a \mid s) \propto \exp(Q(s,a)/\alpha)$. Viewing the KL divergence $\mathrm{KL}(\mu\|\pi_Q^*)$ as a free-energy functional over per-state action distributions μ, one can apply the JKO scheme (see (13) in Section 4.4) to obtain discrete-time W_2 gradient flows. Because the resulting proximal steps are intractable in closed form, the distribution at each iteration is approximated by N particles $\{a_i^{(l)}\}_{i=1}^N$, updated via an SVGD-style kernel rule:

$$a_i^{(l+1)} = a_i^{(l)} + \varepsilon\,\frac{1}{N}\sum_{j=1}^{N}\Big[k(a_j^{(l)}, a_i^{(l)})\,\nabla_{a_j}\tfrac{Q(s,a_j^{(l)})}{\alpha} \;+\; \nabla_{a_j}k(a_j^{(l)}, a_i^{(l)})\Big],$$

where $k(\cdot,\cdot)$ is a positive-definite kernel (Xu et al., 2026). The first term steers particles toward high-Q actions; the second is a repulsive term that maintains diversity. The temperature α controls the exploration–exploitation balance. OT enters here not as a constraint but as the *geometry of the update*: the JKO proximal step moves per-state action distributions along the W_2 gradient flow, with no trajectory-level temporal structure.

Partial OT for suboptimal demonstrations (PPL). Expert data can be expensive, noisy, or simply suboptimal. In that setting, forcing the learner to match all observed behaviour is undesirable. Asadulaev et al. (2024) therefore formulate Partial Policy Learning (PPL) as a partial OT problem: the learner should match only the useful part of the behaviour distribution, identified through Q-values, rather than clone the entire dataset. The intuitive picture is that some of the dataset mass should be copied and some should be ignored. Regarding the policy as a map moving mass from the state distribution $p(s)$ to the action distribution induced by the behaviour policy $\pi_\beta(\cdot \mid s)$, the goal is to learn a policy π that transports only

part of the mass while still favouring high-value actions:

$$\min_{\pi_{\#}p(s)\leq w(\pi_\beta(\cdot|s))} \mathbb{E}_{s\sim\mathcal{D},a\sim\pi(s)}[-Q^\pi(s,a)],$$

where w controls the degree of partiality. When $w = 1$, the problem reduces to standard OT and the learner is pushed toward cloning the behaviour policy with negative Q-value as the cost. When $w > 1$, only part of the behaviour distribution needs to be matched, which creates room to discard low-quality behaviour and improve on the dataset. This flexibility is useful only when the data are genuinely suboptimal: if π_β is already near-expert, taking w too large needlessly loosens the constraint. The resulting partial OT problem is solved through a dual maximin formulation (Gazdieva et al., 2023),

$$\max_{f\geq 0}\min_{\pi} \mathbb{E}_{s\sim\mathcal{D},a\sim\pi(\cdot|s)}[-Q^\pi(s,a) - f(s,a)] + \omega\mathbb{E}_{(s,a)\sim\mathcal{D}}[f(s,a)].$$

with the potential f, policy π, and value function Q all parameterised by neural networks in practice.

Unbalanced OT for constrained policy optimisation. Partial OT relaxes matching by transporting only part of the mass. Unbalanced OT relaxes it differently: mass may be created or destroyed, but doing so incurs a penalty. This is useful when the right policy should not merely select a subset of the data, but should also be allowed to reweight it more flexibly. Givchi et al. (2021) cast policy optimisation in exactly this way, as occupancy-measure matching with negative reward as the transport cost and Bregman penalties, KL in their implementation, on the mismatch between projected marginals. With baseline policy π' representing either the previous iterate or an expert policy, the distributionally constrained objective is

$$\max_{\rho_\pi} \mathbb{E}_\pi[r] - \epsilon_1 D_{\psi_1}(\rho_\pi\mathbf{1} \mid \rho_{\pi'}\mathbf{1}) - \epsilon_2 D_{\psi_2}(\rho_\pi^T\mathbf{1} \mid \rho_{\pi'}^T\mathbf{1}),$$

where $\rho_\pi\mathbf{1}$ represents the projected state marginal and $\rho_\pi^T\mathbf{1}$ represents the projected action marginal of the occupancy measure. Assuming that the set of discounted occupancy measures is a compact convex polytope and adding an entropy term, Givchi et al. (2021) optimise the resulting problem with Dykstra-style Bregman projections (Dykstra, 1983). Considering large or continuous state and action spaces, this is extended to off-policy optimisation with sample-based estimates and duality. While introduced in an online setting, the formulation is directly applicable to offline constraints when data coverage is partial or when strict matching to empirical marginals is undesirable. From a computational standpoint, Fatras et al. (2021) show that minibatch estimation interacts favourably with unbalanced OT: unlike balanced OT, where subsampling introduces bias and artefacts, unbalanced losses with minibatches yield significantly more robust gradient estimates, supporting the scalability of such formulations.

Across all of these methods, the ground cost $c(a, a')$ determines what closeness means in action space. In continuous control a scaled Euclidean cost $c(a, a') = \|a - a'\|_2^2/d$ (with d the action dimension) is a natural default, used by both Q-DOT and BRAC-W. For structured or hybrid action spaces, the cost should encode task-relevant semantics. For instance, grouping actuators by body part or penalising sign reversals more heavily than magnitude changes. The regularisation strength α in (16) must be tuned jointly with critic learning: overly strong constraints prevent improvement over π_β, while overly weak ones reintroduce extrapolation errors. When the behaviour support is narrow, transport outside the dataset support can be unsafe even if the geometric cost is small.

6.3 Summary

OT-based offline RL methods can be read as making the "stay close to data" principle *geometric*. The distance between the learned policy and the behaviour policy is not only a matter of likelihood ratios (KL), but of how much probability mass must be transported in the action space under a chosen cost. Partial and unbalanced OT extend this idea to settings where the dataset contains suboptimal behaviour or outliers, relaxing mass conservation so that the learner matches only the best-supported portion of behaviour.

Several OT tools discussed elsewhere in this survey suggest extensions in offline RL. OT-based transition ambiguity sets (Section 7.1) and OT-based world models (Section 5) can be combined with offline learning to

Figure 11: OT in robustness, safety, and transfer (Section 7). A Wasserstein ambiguity set $\mathcal{U}_\varepsilon$ around the nominal transitions defines worst-case Bellman updates (Section 7.1). OT-based visitation regularisation and preference alignment shape where the policy visits or how it aligns with human feedback (Sections 7.2–7.4).

provide robust/pessimistic targets under model misspecification. OT-based data filtering and reweighting (e.g., through OTR-style scores in Section 7.6) provide a complementary mechanism for reducing distribution shift by curating the dataset before policy learning. Wasserstein/Sinkhorn penalties can also be applied to learned behavioural embeddings, yielding offline analogues of behaviour-guided regularisation (Section 4.3). Establishing systematic guidelines for choosing costs and partiality/unbalanced parameters from data, and providing theory comparable to KL-regularised offline frameworks, remain open directions.

7 Robustness, safety, and transfer under distribution shift

Robustness and safety become central once a learned policy is deployed outside the distribution on which it was trained (Figure 11). Risk criteria such as variance, entropic risk, or CVaR summarise different aspects of return uncertainty, but they say little about *how* uncertainty enters through the dynamics or *which* perturbations should count as plausible. OT gives a geometric alternative: uncertainty is represented as bounded transport of probability mass under a chosen ground metric, and robustness is expressed through worst-case or stay-close objectives in that geometry. The central question is which environment shifts should be treated as plausible during training.

This geometric viewpoint appears in several distinct places. It gives Wasserstein ambiguity sets for distributionally robust control, visitation-level penalties for safety shaping, uncertainty scores derived from transport between value-induced action distributions, DRO formulations for preference shift in RLHF, and alignment tools for transfer across tasks or domains. OT-based curriculum generation via distributional interpolation is discussed separately in Section 8.3.

7.1 Distributionally robust control with Wasserstein ambiguity sets

A standard distributionally robust RL perspective starts from a nominal transition law $P_0(\cdot \mid s,a)$, estimated from data or provided by a simulator, and then asks what happens if the true environment is nearby but not identical. Wasserstein ambiguity sets answer that question geometrically: for each (s,a), nature may replace $P_0(\cdot \mid s,a)$ by any $P(\cdot \mid s,a)$ lying inside a Wasserstein ball of radius ε under the chosen metric. In plain language, the agent is trained not against one model of the world, but against all models that can be reached from the nominal one by a bounded amount of transport.

A canonical formulation (Figure 12) introduces, for each (s,a), a Wasserstein ball around the nominal next-state distribution:

$$\mathcal{U}_\delta(s,a) = \Big\{ P(\cdot \mid s,a) \in \mathcal{P}(\mathcal{S}) : \ \mathsf{W}_p\big(P(\cdot \mid s,a),\, P_0(\cdot \mid s,a)\big) \leq \delta \Big\},$$

and collects these into an ambiguity set over transition kernels,

$$\mathcal{U}_\delta = \Big\{ P : \ P(\cdot \mid s,a) \in \mathcal{U}_\delta(s,a) \ \ \forall (s,a) \in \mathcal{S} \times \mathcal{A} \Big\}.$$

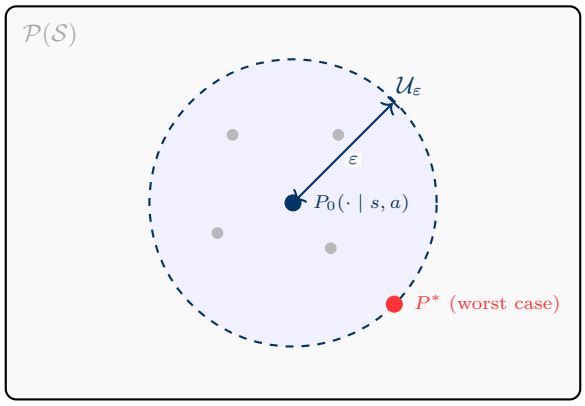

Figure 12: The Wasserstein ambiguity set for distributionally robust RL. The nominal transition $P_0(\cdot \mid s, a)$ (dark-blue dot) sits at the centre of a Wasserstein ball $\mathcal{U}_\delta = \{P : \mathsf{W}_p(P, P_0) \leq \delta\}$ inside the space of transition distributions $\mathcal{P}(\mathcal{S})$. Grey dots denote other candidate distributions inside the ball. The worst-case is represented by P^* (red dot). Two modelling choices shape this game: the ground metric d determines *which* perturbations are considered admissible (through the W_p geometry), while the radius δ controls *how far* models may move away from the nominal one, determining the conservatism of the resulting policy.

The distributionally robust control problem then replaces the Bellman expectation under P_0 with a worst-case expectation over $P \in \mathcal{U}_\delta$:

$$V^\star(s) = \max_{a \in \mathcal{A}} \min_{P(\cdot|s,a) \in \mathcal{U}_\delta(s,a)} \Big\{ R(s,a) + \gamma\, \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^\star(s')] \Big\}.$$

The resulting control problem is minimax: the agent chooses actions to maximise return, while nature chooses the most damaging transition law inside the Wasserstein ball. The radius ε controls how pessimistic this game becomes, and the ground metric determines which perturbations count as nearby enough to be adversarially relevant. Yang (2017) show how Kantorovich duality turns these robust Bellman updates into a finite-dimensional convex program when the nominal distribution has finite support, thereby avoiding direct optimisation over an infinite-dimensional ambiguity set. Clement & Kroer (2021) then develop first-order methods aimed at better scalability.

One concrete example is WR^{2}L (Abdullah et al., 2019), which imposes an expected Wasserstein constraint around a reference transition model $P_0(\cdot \mid s, a)$:

$$\max_\theta \left[\min_\phi \; \mathbb{E}_{\tau \sim p_\theta^\phi(\tau)} \Big[\sum_{t=0}^{T-1} \gamma^t R(s_t, a_t) \Big] \right] \quad \text{s.t.} \quad \mathbb{E}_{(s,a) \sim d_{\pi,\phi}} \big[\mathsf{W}_2^2 \big(P_\phi(\cdot \mid s, a),\, P_0(\cdot \mid s, a) \big) \big] \leq \delta,$$

where the trajectory density $p_\theta^\phi(\tau)$ and state-action visitation distribution $d_{\pi,\phi}$ are induced by policy π_θ under transition model P_ϕ. They approximate the W_2 term using curvature information (via a Hessian-of-expectation approximation) and propose a corresponding optimisation procedure. Conceptually, this treats robustness as a *budgeted* transport of mass in next-state space, with the metric encoding which perturbations are plausible.

The ambiguity-set idea extends naturally to stochastic control with uncertain exogenous disturbances, where duality-based reformulations again yield tractable Bellman operators with out-of-sample guarantees (Yang, 2017; 2020; Kordabad et al., 2022; Kandel & Moura, 2020; Hou et al., 2020; Queeney et al., 2024). More general distributionally robust stochastic optimisation results based on Wasserstein distance are established by Gao & Kleywegt (2023).

7.2 Safety control via visitation regularisation

A complementary approach targets safety more directly by shaping *where* the policy visits. Let ρ_{safe} encode safety preferences (e.g., low mass near obstacles or high mass around safe corridors). One can then incorporate

a transport-based penalty that encourages the policy to move its induced state visitation distribution toward (or away from) ρ_{safe} in the geometry induced by a chosen cost on states. Here OT is not modelling uncertainty in the environment; it is modelling how far the policy's visitaiton distribution is from a safer one.

In discrete state and action settings, this idea can be implemented as a regularisation term in value-based updates. Baheri (2023a) modify the standard Q-learning update by subtracting a state-level OT cost that measures how far the policy's visitation distribution has drifted from the safety prior:

$$Q(s,a) \leftarrow Q(s,a) + \alpha\big[R(s,a) - \lambda\, C(s,s') + \gamma \max_{a'} Q(s',a') - Q(s,a)\big],$$

where $C(s,s') = \Gamma^*_{s,s'}\|s-s'\|^2$, with Γ^* the OT plan for $\mathsf{W}_2^2(\rho_\pi, \rho_{\text{safe}})$, is the optimal transport cost from state s in the policy-induced state visitation ρ_π to state s' in the risk distribution ρ_{safe}, and $\lambda > 0$ tunes the safety–return trade-off. Because $C(s,s')$ is *subtracted* from the update, states that push the visitation far from ρ_{safe} receive a pessimistic Q-value; follow-up experiments in Shahrooei & Baheri (2024) instantiate the same pattern with a squared-Euclidean cost and report practical safety gains on small discrete tasks. This illustrates a common pattern in OT-based safety: rather than imposing hard constraints, one introduces a geometry-aware *soft* mechanism that shifts probability mass away from risky decisions.

This only works as well as the safety prior ρ_{safe} is specified: if that prior is poor, the OT penalty can push learning in the wrong direction.

7.3 Action uncertainty scores via entropic OT

The visitation-based methods above assume a pre-specified risk distribution P_r. When such a prior is unavailable, one can instead read safety off the *learning dynamics themselves*. Safety constraints can also be enforced implicitly by discouraging actions with unpredictable outcomes. Without requiring a reference risk distribution, Shahrooei & Baheri (2025) quantify the uncertainty of discrete actions $\{a_i\}_{i=1}^N$ and encourage the agent to select actions with more predictable outcomes. The total uncertainty in a state s is measured by a Wasserstein distance between action distributions derived from current and target Q-value estimates. The intuition is to look at how much the value-induced distribution over actions changes from one update to the next: if the mass has to be rearranged heavily, then the agent is still unsure about which actions should be preferred. For each state s, they define the uncertainty score

$$U(s,a_i) = \frac{\Delta(s,a_i)}{W_p(Q_s, T_t)}$$

where $Q_s = \sum_{i=1}^N q_i^s \delta_{a_i}$ and $T_t = \sum_{i=1}^N p_i^s \delta_{a_i}$ are two discrete distributions over the available actions. The first is induced by the current Q-values at state s, and the second by the SARSA target values after the latest transition. The OT coupling between these two action distributions is computed with Sinkhorn iterations. The term $\Delta(s,a_i)$ measures the transport imbalance associated with action a_i, namely how much mass linked to that action must be reassigned when passing from the current value profile to the target one. This uncertainty score is used to define the behaviour policy

$$\pi^b(a \mid s) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{|\mathcal{A}|} & \text{if } a = \operatorname{argmax}_{a'}[Q(s,a') - \beta U(s,a')] \\ \frac{\epsilon}{|\mathcal{A}|} & \text{otherwise.} \end{cases}$$

This provides an alternative route to risk-aware behaviour: instead of specifying a safe target distribution explicitly, the method penalises actions whose value estimates move erratically from one update to the next, with the strength controlled by β. At a high level, the method is using OT as a detector of value instability across actions.

7.4 Preference shift and Wasserstein DRO in RLHF

So far, P_0 has always been a transition law or a visitation distribution. The same ambiguity-set template transfers directly to the setting where the uncertain object is instead a *preference distribution* from human

annotators. Wasserstein ambiguity sets also arise in preference-based learning and RLHF for large language model (LLM) alignment (cf. the regularisation perspective in Section 4.2), where the deployed preference distribution may differ from the training distribution due to annotator heterogeneity, selection bias, or temporal drift. In such settings, distributionally robust optimisation (DRO) formulations treat the empirical preference distribution as nominal and optimise for worst-case preference shifts within an ambiguity set, including p-Wasserstein balls. The conceptual analogy with robust control is direct: instead of hedging against shifted transition laws, one now hedges against shifted preference data.

Concretely, Xu et al. (2025) extend Direct Preference Optimisation (DPO) (Rafailov et al., 2023) and Sahu & Wells (2025) extend REBEL (Gao et al., 2024) to a DRO objective of the form

$$\min_{\theta} \max_{\mathbb{P} \in \mathcal{B}_\delta(\mathbb{P}^0; D)} \mathbb{E}_{z \sim \mathbb{P}}[l_{\mathrm{DPO/REBEL}}(z; \theta)],$$

where $\mathbb{P}^0$ is the nominal data-generating distribution, $\mathcal{B}_\delta(\mathbb{P}^0; D)$ is an ambiguity set defined by a distance D (including Wasserstein choices), and $l(\cdot; \theta)$ is the preference-learning loss. From the OT perspective, this matches the same robustness template used for transition uncertainty (Section 7.1): the modelling choices are the geometry defining plausible distribution shifts and the radius ϵ controlling conservatism.

Beyond DRO, several recent works use OT more *directly* as an alignment operator on preference or reward distributions. Distributional preference alignment via optimal transport formulates alignment updates by transporting between predicted and target preference/reward distributions rather than only constraining worst-case shifts (Melnyk et al., 2024). Another fine-grained use of OT is at the *token level*: OT couplings between token representations can be used to assign saliency/importance weights within preference pairs, biasing the optimisation toward the parts of responses that drive preference differences (Li et al., 2025). Finally, dual-reference "push–pull" distributional alignment uses two reference distributions (desired and undesired) and OT-based weighting to push the model toward the former while pulling it away from the latter during fine-tuning (Wang et al., 2026).

While these works are not RL benchmarks in the classical sense, they are directly relevant to preference-based sequential decision-making because they provide concrete recipes for (i) modelling preference shift geometrically and (ii) propagating sparse feedback through an embedding space. An important open direction is how these OT-based alignment mechanisms extend to temporally extended preferences and trajectory-level feedback, where preferences are defined over sequences of states/actions rather than i.i.d. samples.

7.5 Reward ambiguity and Wasserstein structure

The preceding subsection hedged against uncertain preference *data*. A closely related source of uncertainty lives one level up, in the reward function itself: even with clean preference data, many reward functions can explain the same observed behaviour. A central difficulty in IRL is therefore *reward ambiguity*: many distinct reward (or cost) functions can induce the same expert policy (or an observationally indistinguishable occupancy). Baheri (2023b) proposes to study this ambiguity through a geometric lens by equipping the set of rewards consistent with the expert with a Wasserstein metric, so that ambiguity can be quantified by distances and dispersion in a transport geometry.

Let $\mathcal{R}(\pi_E)$ denote a set of reward functions consistent with the expert behaviour. The key modelling step is a representation map (denoted Φ in (Baheri, 2023b); we write Ξ here to avoid clashing with the survey-wide use of Φ for behavioural embeddings)

$$\Xi : \mathcal{R}(\pi_E) \to \mathcal{P}(\mathcal{X}), \qquad R \mapsto \mu_R,$$

which embeds each candidate reward as a probability measure on the Wasserstein space $\mathcal{X}$ equipped with a ground metric. Given candidates $\{R_i\}_{i=1}^N \subset \mathcal{R}(\pi_E)$, a "centroid" reward function representation can be identified as

$$\eta^\star = \operatorname*{argmin}_{\eta \in \mathcal{P}(\mathcal{X})} \sum_{i=1}^{N} \mathsf{W}_p\big(\eta, \mu_{R_i}\big).$$

The average pairwise distance $\frac{2}{N(N-1)}\sum_{i<j}\mathsf{W}_p(\mu_{R_i},\mu_{R_j})$ provides a quantitative notion of "how ambiguous" the rewards are under the chosen representation and ground metric, which shows the increasing importance of identifying ambiguity with the higher dimension of state-action space (Baheri, 2023b).

This use of OT is conceptually orthogonal to the behaviour-matching pipelines in Section 3: here OT characterises *reward-space structure*, whereas those methods use OT directly on *behaviour distributions* (occupancies, trajectories, or embeddings). The notion of reward ambiguity connects directly to the Wasserstein ambiguity sets used for transition uncertainty (Section 7.1) and preference shifts in RLHF (Section 7.4), though the latter concern transitions or preferences rather than reward functions. In practice, conclusions depend strongly on the choice of Ξ and the metric on $\mathcal{X}$, which determine whether Wasserstein distances capture control-relevant modes of ambiguity in high-dimensional reward families (Baheri, 2023b). This ambiguity measure is representation-dependent and is not invariant to reward shaping unless Ξ encodes shaping equivalence.

7.6 Transfer

Beyond robustness and safety, OT also appears in transfer across tasks and domains. Here the goal is no longer to hedge against worst-case perturbations, but to align source and target distributions so that policies, rewards, or representations transfer meaningfully under shift. The common thread is still geometric correspondence: OT specifies which source states, trajectories, or observations should count as close to which target ones. Robustness asks which shifts should be feared; transfer asks which correspondences should be trusted.

Goal/intent alignment. Zero-shot offline imitation (Pathak et al., 2018) aims to pretrain a task-agnostic policy and then adapt it to unseen tasks from only a single demonstration. Rupf et al. (2024) do this by aligning state occupancies to a sequence of goals with OT. With the training set

$$\mathcal{D} = \{(s_0, a_0, s_1, a_1, \dots)\}$$

and a single rough (ordered but without time-step information) or partial (partial information is contained in each state) expert demonstration

$$\mathcal{D}^e = (g_0, g_1, \dots, g_M),$$

where the goal g is an abstract representation of state s, the optimisation problem is

$$\pi^* = \operatorname*{argmin}_{\pi} \mathrm{OT}_c(\hat{\rho}^{\pi}_N, \hat{\rho}^E_M),$$

where $c(s,g) = V^*(s,g)$ is a goal-conditioned value function trained from $\mathcal{D}$, and $\hat{\rho}^{\pi}_N, \hat{\rho}^E_M$ are empirical estimates of state and goal occupancies. The advantage of this formulation is that it aligns full occupancy patterns rather than greedily matching immediate goals, which reduces the myopia of earlier goal-conditioned methods. The paper also tests unbalanced OT, but does not find a significant gain in that setting.

Preference transfer. Preference-based RL addresses reward design by learning from human comparisons between trajectories, but collecting such preferences for every new task is costly. Liu et al. (2024) therefore study zero-shot preference transfer. When tasks share an action space, they solve a GW problem between source and target trajectory sets and use the resulting alignment to transfer preference labels from source-task trajectory pairs to target-task pairs. GW is natural here because the state observations can differ across tasks, while their intra-trajectory relations may still be comparable.

Let us assume that there are trajectories $\{x_i\}_{i=1}^M$ for the source task with pairwise preferences $z(x_i, x_{i'})$ and trajectories $\{y_j\}_{j=1}^N$ for the target task . With weights $\sum_{i=1}^M u_i = 1$ and $\sum_{j=1}^N v_j = 1$, let Γ be the optimal transport matrix for GW between two probability measures $\mu = \sum_{i=1}^M u_i \delta_{x_i}$ and $\nu = \sum_{j=1}^N v_j \delta_{y_j}$. PEARL (Liu et al., 2024) defines the matching matrix for each pair of trajectories $(y_j, y_{j'})$ in the target task as $A^{jj'} = \Gamma_{\cdot j}\Gamma^T_{\cdot j'}$, where $\Gamma_{\cdot j}$ is the vector of transported mass from all source trajectories into y_j. The target trajectories are then labelled with preferences

$$z(y_j, y_{j'}) = \sum_i \sum_{i' \neq i} A^{jj'}_{ii'} z(x_i, x_{i'}).$$

Thus source pairs that align more strongly with a target pair contribute more heavily to the transferred label.

Visual/domain adaptation. In visual domains, the main challenge is to learn a representation that removes domain-specific appearance while preserving task-relevant content. Building on PPO and borrowing the critic-based Wasserstein idea from WGAN, WAPPO (Roy & Konidaris, 2021) learns a representation function h_θ that approximates the optimal one by minimising the W_1 distance between source and target feature distributions,

$$\mathsf{W}_1((h_\theta)_\# P_s, (h_\theta)_\# P_t) = \sup_{\|f\|_{\mathrm{Lip}} \le 1} \{\mathbb{E}_{x\sim P_s} f(h_\theta(x)) - \mathbb{E}_{x\sim P_t} f(h_\theta(x))\}$$

where f is the adversarial critic (1-Lipschitz potential) trained to approximate the supremum. Similarly, WDGRL (Shen et al., 2018) adapts data to different domains by learning a shared feature space together with a task classifier on labelled source data. The domain discrepancy is measured by a Wasserstein critic in dual form, with the 1-Lipschitz constraint enforced by a gradient penalty.

Unbalanced minibatch OT has also been applied for domain adaptation to learn the representation function and the label function for the target domain (Fatras et al., 2021) as mentioned in Section 6.2. Moreover, a Wasserstein-based reward has been proposed to select intermediate domains in continuous domain adaptation (Liu et al., 2025).

Data filtering via OT scores. When a target domain has only limited offline data, OT-based alignment can identify which source-domain transitions are most compatible with the target dynamics before policy learning. OTDF (Lyu et al., 2025) instantiates this in cross-domain offline policy adaptation by solving OT between empirical measures over transition tuples $u = (s, a, s')$ from a source dataset $\mathcal{D}_{\mathrm{src}}$ and $u' = (\bar{s}, \bar{a}, \bar{s}')$ from a (small) target dataset $\mathcal{D}_{\mathrm{tar}}$, using a cost $C(u, u')$ on $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$. The optimal coupling Γ^* yields a per-transition compatibility score (an expected cost under Γ^*), which is then used to filter or weight source transitions during critic learning. OTDF further adds an explicit dataset-constraint regulariser that encourages the learned policy to stay within the support of the target dataset, reducing bias toward the source behaviour when $\mathcal{D}_{\mathrm{tar}}$ is small (Lyu et al., 2025). OT has also been explored in federated IRL settings, where expert demonstrations are distributed across clients and OT aggregation can improve reward recovery under heterogeneity (Millard & Baheri, 2026).

7.7 Summary

Across the applications in this section, OT turns robustness and transfer questions into geometric ones. Ambiguity sets ask which transition or preference shifts should count as plausible. Visitation shaping asks how far the policy's visitation distribution is from a safer target distribution. Action-uncertainty scores ask which actions become unstable when value-induced transport changes sharply. Transfer reuses the same logic to align source and target distributions across tasks or domains. In every case, the decisive modelling choices are the ground metric, which defines what kinds of shifts are plausible, and the scale parameter, such as ε, λ or β, which determines how conservative the method becomes. These methods fail when the ball radius, safety prior, or feature metric is poorly calibrated.

8 Other applications

OT appears in RL well beyond the main themes of imitation, policy optimisation, world modelling, offline learning, and robustness. This section gathers those smaller but still instructive pockets of the literature: diversity and skill discovery, multi-agent coordination, curricula, and a handful of generative or evaluation settings. What links them is not a single algorithmic template, but a recurring move: replace an ad hoc similarity or diversity measure by a transport-based one on policy-induced distributions, rollout-derived empirical measures, or learned embeddings. The same diagnostic question applies throughout: which distributions are being compared, and what does the transport geometry add? This section is best read as frontier and adjacent applications; the evidence base is less mature than in imitation, offline RL, and policy optimisation. When a paper uses a bespoke distance denoted "WD", we interpret it according to the paper's definition as a Wasserstein distance W_p or a Sinkhorn-type surrogate.

8.1 Multiple policies or skills

In unsupervised RL and skill discovery, the problem is often not how to imitate a target behaviour, but how to make several behaviours genuinely different from one another. OT offers a direct way to do this by separating the induced state distributions p_l of concurrently trained policies. Here the transport distance itself becomes the diversity measure. He et al. (2022) propose an intrinsic reward based on projected Wasserstein discrepancies (Rowland et al., 2019) between state marginals, so that each policy is pushed away from its closest neighbour and state-space coverage improves. He et al. (2022) define the final reward of an episode induced by policy l as $\min_{l'=1,l\neq l}^{N} \mathsf{W}(p_l, p_{l'})$ in practice, while the amortised reward for a single state s_i is assigned as

$$r(s_i) = \sum_{k=1}^{m} \sum_{j=1}^{n} c(s_j, s'_k)\Gamma^*_{i,k},$$

where Γ^* is the optimal transport matrix between two state distributions. The global expected cost is used here, comparing with the local reward in OTR. Projecting states through a representation f mitigates the curse of dimensionality and makes costs task-aware. The practical interpretation is that a skill gets rewarded for visiting regions of state space that the other skills have not already claimed.

One can push the same idea further without explicitly constructing a per-step reward. Yang et al. (2025) use the Wasserstein distance to improve mutual-information skill learning by enforcing diversity and separability directly at the level of state distributions. Their objective (WSEP) maximises

$$\text{WSEP} = \sum_{z_i \in \mathcal{Z}} \sum_{z_j \in \mathcal{Z}\setminus\{z_i\}} \mathsf{W}(p(S \mid z_i), p(S \mid z_j)),$$

where $p(S \mid z)$ is the state distribution of a policy, and $\mathcal{Z}$ indexes skills. They further show that vertices of the feasible state-distribution polytope can be discovered via iterations of maximising projected WSEP,

$$\text{PWSEP}(i) = \min_{\lambda} \mathsf{W}\Big(p(S \mid z_i), \sum_{z_j \in \mathcal{Z}_i} \lambda^j p(S \mid z_j)\Big),$$

where at iteration i, $\mathcal{Z}_i$ denotes previously learned skills and λ are mixture coefficients with $\sum_{j\in\{0,1,\ldots,i-1\}} \lambda^j = 1$. In practice, the sliced Wasserstein distance (Kolouri et al., 2019) improves computational efficiency. This can be read as a "novelty against mixtures" criterion: a new skill should not just differ from one old skill, but from combinations of the old ones.

Quality-diversity imitation learning aims to learn diverse behaviours from limited demonstrations. WQDIL (Yu et al., 2024) improves reward-learning stability via Wasserstein adversarial training within the latent space of a Wasserstein auto-encoder (WAE). For WAE-WGAN, the objective is

$$D_{\text{WAE}}(P_X, P_G) = \inf_{Q(z|x)\in\mathcal{Q}} \mathbb{E}_{P_X} \mathbb{E}_{Q(z|x)}[c(x, P_G(z))] + \lambda \mathsf{W}_1(Q_z, P_z),$$

where P_X is the data distribution, P_G the generative model, Q_z the distribution of encoded demonstrations and P_z the prior. The first term is related to W_2^2 with c as the squared Euclidean cost. The corresponding loss for WAE-WGAIL is

$$\begin{aligned}\mathcal{L}(\phi, \psi) =& \mathbb{E}_{x^e\sim\mathcal{D}, z^e\sim Q_\phi(z^e|x^e)}[c(x^e, G_\psi(z^e))] + \mathbb{E}_{x^\pi\sim\pi, z^\pi\sim Q_\phi(z^\pi|x^\pi)}[c(x^\pi, G_\psi(z^\pi))] \\ &+ \lambda\, \mathbb{E}_{z^e, z^\pi}[D_{\mathsf{W}}(z^e) - D_{\mathsf{W}}(z^\pi)],\end{aligned}$$

where Q_ϕ is the encoder, G_ψ the decoder, D_{W} the 1-Lipschitz potential function for the dual form of W_1, and $x = (s, a, \delta(s))$.

Wasserstein distances are also applied to hierarchical RL to improve diversity among subpolicies (Li et al., 2023). For each subpolicy π_k, a regularisation term is added to the actor loss:

$$\mathcal{L}(\pi_k) = -\min_{j\neq k} \mathsf{W}_\varepsilon(\mu_{\pi_k}, \mu_{\pi_j}),$$

where policies are represented as categorical distributions for discrete actions and Gaussians for continuous actions, with behavioural embeddings analogous to Pacchiano et al. (2020) (cf. Section 4.3). HiPBOT (Le et al., 2023) formulates blending multiple expert policies as an entropy-regularised unbalanced OT problem to derive mixture weights.

8.2 Multi-agent RL

Multi-agent RL extends the single-agent setting to environments where several agents interact, each with its own reward function. A convenient formalism is a discounted Markov game

$$\mathcal{G} = \big(\mathcal{S}, \{\mathcal{A}_i\}_{i=1}^N, P, \{R_i\}_{i=1}^N, \gamma, p_0\big),$$

with joint action space $\mathcal{A} = \prod_{i=1}^N \mathcal{A}_i$ and joint action $a = (a^1, \dots, a^N)$. A (Markov) joint policy factorises as $\pi(a \mid s) = \prod_{i=1}^N \pi_i(a^i \mid s)$ in the fully observed case (or $\pi_i(a^i \mid o^i)$ under partial observability). The joint discounted occupancy measure follows the single-agent definition (3),

$$\rho_\pi(s, a) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \, \Pr(s_t = s, a_t = a \mid \pi),$$

and per-agent marginals (useful when agents have local actions) are $\rho_\pi^i(s, a^i) = \sum_{a^{-i}} \rho_\pi(s, a)$.

Where OT can enter MARL objectives. Many OT ideas from the earlier sections extend naturally to MARL, for example trust regions on per-agent policies, Sinkhorn-regularised behaviour distributions, or visitation-shaping objectives. What changes is the object to which OT is applied. One must decide whether to regularise the joint occupancy ρ_π, per-agent marginals ρ_π^i, or some learned multi-agent embedding, and that choice determines both the geometry and the computational burden. This is the main conceptual fork in MARL: does OT compare the behaviour of the team as a whole, or the behaviour of each agent separately? Baheri & Kochenderfer (2024) survey several such roles for OT in cooperative MARL, from aligning agent strategies to a collective objective, to treating resource allocation as a transport problem, to using Wasserstein drift between state distributions as a signal of nonstationarity.

Consensus via Sinkhorn barycentres. Coordination can be posed directly in distribution space rather than parameter space. Baheri (2025) compute a Sinkhorn-divergence barycentre and regularise each agent toward it, encouraging consensus in visitation distributions while maintaining decentralised policies. A generic template consistent with this description is

$$\bar{\rho} \in \operatorname*{argmin}_{\rho} \sum_{i=1}^{N} \mathsf{S}_\varepsilon\big(\rho_{\pi_i}, \rho\big), \qquad \pi_i \leftarrow \operatorname*{argmax}_{\pi_i} \; \mathbb{E}_{\rho_\pi}[R_{\text{common}}] \; - \; \lambda \, \mathsf{S}_\varepsilon\big(\rho_{\pi_i}, \bar{\rho}\big),$$

where ρ_{π_i} is the (local or joint) visitation measure associated with agent i and $\bar{\rho}$ is the Sinkhorn barycentre acting as a consensus target. The barycentre plays the role of a soft team average, and each agent is nudged toward that shared behavioural centre.

When the number of agents is large, mean-field formulations replace the joint interaction by a population distribution and can make OT geometry explicit: the population distribution becomes the evolving object, and Wasserstein/OT structure can be used to describe (and sometimes stabilise) its dynamics. Recent mean-field actor–critic flow approaches learn mean-field games by explicitly coupling policy/value updates to an evolution of the population distribution in a Wasserstein geometry, providing a concrete instance of "learning dynamics over distributions" beyond single-agent RL (Zhou et al., 2025). This line aligns with the survey's gradient-flow viewpoint (Section 4.4), but with the population distribution playing the role of the transported measure. The Wasserstein distance can also express robustness in many-agent settings: ERNIE (Bukharin et al., 2023) proposes adopting Wasserstein constraints in a mean-field approximation to handle many-agent uncertainty, turning nonstationarity into a distributional uncertainty set over population measures and suggesting a natural bridge to the ambiguity-set constructions of Section 7.1. A closely related federated-RL setting uses Wasserstein barycentres to aggregate heterogeneous client policies (Pereira & Amini, 2025).

8.3 Curricula via distributional interpolation or constraints

OT can generate intermediate tasks that bridge an easy source to a difficult target, replacing direct transfer with a graduated curriculum. Here OT is being used to answer a curriculum-design question: which task should come next, given both where the target lies and what the current policy can already solve?

Contextual RL can be modelled via a contextual MDP $\mathcal{M}(c) = (\mathcal{S}, \mathcal{A}, P_c, R_c, \gamma_c, \rho_{0,c})$, where c is a context variable and the training objective becomes $\mathbb{E}_{\mu(c)}[J(\pi, c)]$, with $\mu(c)$ defining the target task distribution. Curriculum RL generates a sequence of intermediate task distributions p_i designed to improve exploration and final performance on $\mu(c)$. The key tension is that the next task should be close enough to the target to be useful, but not so hard that the current policy cannot learn from it.

Constrained OT with a W_2 proximity constraint generates intermediate task distributions that are close to the target while remaining solvable by the current policy (Klink et al., 2022; 2024):

$$p(c) = \operatorname*{argmin}_{p} \ \mathsf{W}_2(p(c), \mu(c)) \quad \text{s.t.} \quad p(c) > 0 \Rightarrow J(\pi, c) \geq \delta_J \ \ \forall c \in \mathcal{C}, \quad \mathsf{W}_2(p(c), q(c)) \leq \delta,$$

where δ_J is the desired performance threshold under $p(c)$ and δ controls deviation from the previous curriculum $q(c)$.

Huang et al. (2022) propose GRADIENT, which defines a π-contextual distance that accounts for policy performance to interpolate between context distributions while maintaining feasibility; in practice they approximate the contextual metric when exact computation is costly. They also use 1-Wasserstein barycentres between distributions μ_0 and μ_1: $\eta_\alpha = \operatorname{argmin}_\eta (1-\alpha)\mathsf{W}_1(\mu_0, \eta) + \alpha \mathsf{W}_1(\eta, \mu_1)$, where α controls the weights.

Alternative curricula: HGG. In goal-conditioned settings, Hindsight Goal Generation (HGG) (Ren et al., 2019) selects intermediate goals that are closer (in a transport-aware sense) to the target goal while remaining achievable by the current policy, yielding more sample-efficient curricula than simple Hindsight Experience Reply (HER) (Andrychowicz et al., 2017) which relabels transitions with achieved goals to mitigate the challenge of sparse rewards. Let $\mathcal{G}$ be the goal space and $\zeta : \mathcal{S} \times \mathcal{G} \to [0, 1]$ be a joint distribution over the initial state and the goal. The agent seeks to maximise $J(\pi; \zeta^*) = \mathbb{E}_{(s,g)\sim\zeta^*}[J(\pi, s, g)]$. To generate goals ζ, HGG optimises

$$\max_{\pi, \zeta : |\zeta| = K} J(\pi; \zeta) - L \cdot \mathrm{OT}_c(\zeta, \zeta^*) \quad \text{s.t.} \quad \sum_{s_0, s_t \in \tau} \mathbf{1}[(s_0, \phi(s_t)) \in \zeta] \leq 1, \quad \forall \tau \in \mathcal{B},$$
$$\sum_{\tau \in \mathcal{B}} \sum_{s_0, s_t \in \tau} \mathbf{1}[(s_0, \phi(s_t)) \in \zeta] = K,$$

where K trajectories are sampled from the replay buffer $\mathcal{B}$, ϕ maps states to goals, and OT_c is the OT distance with cost

$$c((s, g), (s', g')) = \alpha \|\phi(s) - \phi(s')\|_2 + \|g - g'\|_2,$$

where $\alpha > 0$ is a weight for the trade-off. This optimisation is solved by iterating between improving the policy and updating the hindsight distribution ζ (solving a Wasserstein barycentre problem via bipartite matching). Safety-prioritising curricula extend this framework to constrained RL, generating intermediate task distributions that explicitly balance learning progress against safety constraints (Köprülü et al., 2025).

8.4 Further OT–RL intersections

Several OT uses fall outside the main templates of the survey but are still worth recording. In generative and sequence-based settings, for example, Sinkhorn-based decoders have been used for non-autoregressive trajectory generation in combinatorial domains such as the travelling salesman problem, where Sinkhorn acts as a differentiable approximation to a permutation or assignment operator (Goh et al., 2022). Chen et al. (2020) add a Wasserstein regulariser between generated and reference sequence embeddings in a learned semantic space, and Sun et al. (2025) use OT-guided matching inside a score-based diffusion policy. The common pattern is that OT is not comparing occupancies or Bellman targets directly; it is supplying a structured discrepancy inside a generative model.

OT also provides useful signals for evaluation and exploration. Nkhumise et al. (2025) use the W_1 distance between successive occupancy measures $\mathsf{W}_1(\rho_{\pi_k}, \rho_{\pi_{k+1}})$ to assess the learning efficiency of RL algorithms, defining optimality measures based on the ratio of total transport effort to final policy improvement. As an intrinsic control signal, Durugkar et al. (2021) propose Wasserstein Distance Maximising Intrinsic Control, where each skill is trained to maximise the Wasserstein distance between its state-visitation distribution and its start-state distribution, encouraging controllable, wide-coverage behaviour without hand-crafted reward shaping, complementing the mutual-information skill discovery methods of Section 8.1.

Finally, OT-based robustness and structural tools extend to settings adjacent to standard RL. In contextual bandits, which can be read as a one-step special case of RL, Shen et al. (2024) develop Wasserstein distributionally robust policy evaluation and learning through tractable OT-dual objectives. The same ambiguity-set logic could naturally feed into robust off-policy evaluation in sequential settings. Two further directions show how far OT now reaches beyond the standard RL template: dynamic OT has been used to derive conditions for counterfactual identifiability via continuous-time flows (De Sousa Ribeiro et al., 2025), with possible relevance to causal RL and offline evaluation, and optimal transport has been used to fuse task-specific models by aligning feature-space geometry under OT plans rather than interpolating parameters directly (Pan et al., 2025), enabling continual fusion without revisiting past tasks.

8.5 Summary

Even in these more specialised settings, the same two questions keep reappearing: *what* distributions are being compared, and *what* geometry makes that comparison meaningful? Once those are fixed, the rest of the design usually follows. In multi-agent settings, one further choice becomes unavoidable: should OT act on joint distributions, which capture coordination but scale poorly, or on marginals, which are cheaper but may miss interactions between agents? The methods in this section should be read as promising but less settled: several trade statistical efficiency or wall-clock cost for a more structured notion of diversity, alignment, or evaluation.

9 Discussion

Across Sections 3–8, the same modelling choices and failure modes recur. This section gathers them in one place: we discuss the main trade-offs, practical guidance, and open problems. Method-level details remain in Tables 3–8; the goal here is synthesis rather than cataloguing.

9.1 Cross-cutting trade-offs

The same tensions keep reappearing across applications. They are best read as design trade-offs rather than universal rules, and they cluster around three recurring themes: sharpness versus stability, the role of representation and temporal structure, and the calibration of the hyperparameters that mediate both.

The first tension is between *sharpness* and *stability*. W_1 dual critics remain informative when distributions barely overlap, which is one reason OT is introduced in adversarial imitation, offline behaviour constraints, and robust ambiguity sets. The price is fragility: Lipschitz enforcement is delicate, and critics can overfit small minibatches (Arjovsky et al., 2017; Gulrajani et al., 2017). Entropic OT is smoother, easier to minibatch, computationally efficient, and often more stable, but biased unless debiased through the Sinkhorn divergence S_ε (Genevay et al., 2018). The same tension propagates to the algorithmic level: dual-critic W_1 methods such as WAIL and BRAC-W depend heavily on critic capacity and optimisation balance; map-based W_2 methods avoid adversarial training but impose convexity constraints through ICNN parameterisations (Brenier, 1991; Omura et al., 2025); primal Sinkhorn objectives sidestep both at the price of making the entropic parameter a sensitive modelling choice. OT does not remove the usual optimisation trade-offs; it relocates them into the choice of solver and geometry, and in practice the way OT is *solved* often matters as much as the RL algorithm it is inserted into.

The second tension concerns what the ground cost actually encodes. The effectiveness of OT geometry is primarily determined by the geometry encoded by the ground cost and the learned features when the cost is

defined via a representation space. Task-specific learned features can improve performance in high-dimensional or unstructured observation spaces. For example, (Bobrin et al., 2024) train the temporal goal-conditioned value function through ICVF (Ghosh et al., 2023), and use the backward embedding φ as representation for OT, where $\|\varphi(s_i) - \varphi(s_j)\|^2$ estimates the average path length from state s_i to s_j, which measures the reachability and is robust to noise in the original state space. Pacchiano et al. (2020) instead suggest a Behavioral Embedding Map (BEM) to abstract behaviours from trajectories or states and apply different BEM in different tasks and settings. The representation becomes a critical design choice for optimisation, and the performance may degrade under embedding-distribution shift even when with freezed representations during training. Temporal structure raises a parallel question of granularity: bag-of-states matching is cheap and effective for short, well-aligned episodes (e.g. OTR, Section 3.4), but creates spurious correspondences under speed variation or partial observability. Banded masks and context windows recover local order at moderate cost (Section 3.5), while fully dynamic couplings remain expensive. The same geometry-versus-ordering tension resurfaces wherever OT operates on sequential data: offline reward labelling (Section 6), curricula (Section 8.3), and cross-domain transfer (Section 7.6). Empirical Wasserstein estimates can converge slowly in high dimension, so raw observations are rarely the right transport space. This motivates feature costs, slicing, minibatches, and entropic or unbalanced relaxations, but each changes the geometry being optimised.

Overlaying both is a *calibration* problem that the OT-RL literature largely avoids. The entropic parameter ε, the radius δ in ambiguity sets (Section 7), the transported mass m in partial OT (Section 6), penalty strengths in unbalanced OT, and proximity radius in curricula (Section 8.3) all require calibration, and no principled, data-driven method exists for any of them. The consequences are asymmetric: too tight a radius produces fragility, too loose a radius produces undue pessimism—yet sensitivity to these parameters is seldom reported.

9.2 When OT helps—and when it does not

OT is most useful when distributions overlap weakly, as in early-training imitation or under severe distribution shift; when the state or action space has a geometry that KL divergence cannot express; or when domains are heterogeneous enough that no sensible pointwise cross-domain cost exists and Gromov–Wasserstein alignment becomes the natural alternative (Mémoli, 2011; Fickinger et al., 2022). It is also a natural fit when the algorithm needs an explicit transport plan, for example for reward labelling, data filtering, or transport-map policies, and when targets are multi-modal enough that KL's mode-seeking behaviour becomes unhelpful.

OT is also easy to overuse. When policy and target distributions already overlap well, for example once an imitation policy is already close to the expert, KL- and JS-based gradients are often informative enough and cheaper. When the action space carries no meaningful geometry a KL-based trust-region or trust-region-like updates such as TRPO and PPO often work perfectly well without the extra modelling burden (Schulman et al., 2015; 2017). And if the ground cost does not encode anything task-relevant, the geometric machinery mostly adds overhead. Recent work makes that point explicitly: proximity-based and lightweight temporal baselines can match OT reward labelling on broad benchmarks (Dong et al., 2025), which raises the practical question of when coupling structure is really necessary rather than simply elegant. Computation sharpens the same issue. In very-small-sample regimes, minibatch Wasserstein estimates are biased and high-variance (Fatras et al., 2020; Bellemare et al., 2017), and in compute-limited settings the inner-loop cost of Sinkhorn iterations, critic training, or ICNN updates may not justify itself. OT is least attractive when demonstrations already overlap with the learner, action geometry is uninformative, behaviour support is too narrow to leave safely, or a Wasserstein radius cannot be calibrated.

9.3 Practitioner's guide

Table 1 is meant as a starting point rather than a prescription. It asks a practical question: if the RL problem looks like this, which OT default is the least bad place to begin? The point is not that these defaults are optimal, but that they provide a reasonable first experiment before heavier tuning. These defaults assume a meaningful ground cost and enough samples to estimate the transport signal.

Two practical themes cut across these defaults. The first is that the ground cost deserves as much attention as the OT solver. Task-specific features can induce a more appropriate geometry than raw observations, but

Table 1: Decision guide: recommended OT defaults by application area.

Application	Start with	Distributions	Cost default	Key pitfall
Online imitation	S_ε (Sinkhorn div.)	occupancies ρ_π vs ρ_{π_E}	learned features	ε sensitive to cost scale
Offline reward labelling	entropic OT + mask	trajectory segments	cosine on encoder	spurious matches without temporal mask
Policy trust region	per-state Sinkhorn	$\pi(\cdot\|s)$ vs $\pi_{\text{old}}(\cdot\|s)$	Euclidean (actions)	radius too tight $\to$ no improvement
Offline constraint	W_1 penalty	$\pi(\cdot\|s)$ vs $\pi_\beta(\cdot\|s)$	action-space metric	overly strong $\to$ clones π_β
World model	W_1 or S_ε	$P_\phi(\cdot\|s,a)$ vs $P(\cdot\|s,a)$	latent features	optimising prediction, not control
Robustness	W_p ball	transitions $P(\cdot\|s,a)$ around P_0	env. geometry	uncalibrated $\varepsilon \to$ over-conservatism
Transfer / filtering	OT scores + soft reweight	behaviour embeddings	shared repr. space	hard thresholding with noisy costs
Curriculum	barycentres / OT path	task distributions	task metric	infeasible interpolated tasks

mis-specified features may distort the metric structure and degrade downstream performance. Heterogeneous units (e.g. joint angles and end-effector positions) should be normalised before being combined, and costs should be scaled before setting ε because entropic regularisation is sensitive to magnitude. For action spaces, it is usually more important to encode task semantics and anisotropy than to default to the isotropic Euclidean distance. Gromov–Wasserstein is best reserved for cases where a pointwise cross-domain cost is genuinely ill-posed; otherwise, a shared representation with standard W_p or Sinkhorn objectives is simpler and usually more stable (Shen et al., 2018; Lyu et al., 2025).

The second theme is that most failures in OT-based RL are numerical or representational rather than algorithmic. Sinkhorn objectives benefit from log-domain iterations and from scaling ε relative to a typical pairwise cost, and warm-starting dual variables helps when the compared distributions move only gradually. Dual W_1 critics need careful Lipschitz control through gradient penalties or spectral normalisation, and should not be grossly over-parameterised relative to the minibatch size (Arjovsky et al., 2017; Gulrajani et al., 2017); W_2 and ICNN-based maps require nonnegative weights for convexity and some regularisation for smoothness (Makkuva et al., 2020). At the trajectory level, masks and context windows are usually safer than unconstrained long-horizon couplings, with the window size chosen according to the dynamics of the environment. Minibatch OT should generally be run on features rather than raw observations, and unbalanced losses are often more robust to subsampling than balanced ones (Fatras et al., 2021). When OT appears as an auxiliary term, e.g. an intrinsic reward, diversity bonus, or filtering weight, annealing its contribution over training can be beneficial as coverage improves. When costs are noisy, soft reweighting is safer than hard thresholding.

9.4 Open problems

Four open directions stand out.

When does coupling structure actually matter? As noted in Section 9.2, good geometry can matter more than solving OT: many papers gain as much from representation design as from the transport formulation, and proximity baselines can match OT reward labelling on broad benchmarks (Dong et al., 2025). What is missing are controlled experiments that isolate the transport component from the representation component, and theoretical characterisations of the regimes, including multi-modal demonstrations, systematic mass mismatch, severe temporal misalignment and heterogeneous domains, where coupling structure is genuinely necessary rather than merely elegant.

Scalability and amortisation. Current OT-RL methods invoke transport solvers per update, per state, per agent, or per minibatch. Sliced and projected approximations reduce this cost while retaining meaningful geometry (Kolouri et al., 2019; Mahey et al., 2023; Chapel et al., 2025), and amortised or neural OT methods learn transport maps or potentials once and reuse them across iterations (Geuter et al., 2025; Chen et al., 2025; Makkuva et al., 2020; Korotin et al., 2023). Flow-matching constructions offer a complementary route, estimating dynamics gaps via Wasserstein distances between source and target transitions (Kong et al., 2025). But the deeper challenge is temporal: scaling OT beyond short windows and masks to long-horizon and high-dimensional control remains largely unexplored.

Mass mismatch as a default. Real offline datasets and demonstrations are almost always incomplete, mixed-quality, or multi-domain, so balanced OT is often the wrong model. Partial and unbalanced OT are natural alternatives (Asadulaev et al., 2024; Chizat et al., 2018; Lyu et al., 2025), and minibatch unbalanced OT mitigates some of the bias inherent in minibatch estimation (Fatras et al., 2021). The missing piece is principled calibration: selecting transported mass and marginal-penalty strengths from observable diagnostics, such as coverage statistics, critic uncertainty and held-out value error, with guarantees analogous to those available for KL regularisation.

Theory for OT-regularised updates. Guarantees for Wasserstein trust regions, Wasserstein natural gradients, and OT-based mirror descent remain thin under realistic function approximation (Zhang et al., 2018; Moskovitz et al., 2021; Li & Montúfar, 2018; Pfau et al., 2025). Recent algorithms that place Wasserstein geometry at the core of the update, e.g. Wasserstein Policy Optimisation, make this gap more pressing. Understanding stability, sample complexity, and approximation error for Wasserstein-proximal updates, relative to the mature theory for KL-based TRPO/NPG, likely requires analyses that separate representation error (ground-cost misspecification) from optimisation error (critic, Sinkhorn, or map approximation).

OT has given RL a useful geometric vocabulary for comparing distributions. The next step is to understand when that geometry genuinely changes the problem for the better, and when it mainly adds machinery. Answering that requires better calibration tools, better scalability, and theory that separates the effect of transport geometry from the effect of representation learning or optimisation heuristics.

References

Mohammed Amin Abdullah, Hang Ren, Haitham Bou Ammar, Vladimir Milenkovic, Rui Luo, Mingtian Zhang, and Jun Wang. Wasserstein robust reinforcement learning. *arXiv preprint arXiv:1907.13196*, 2019.

David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. In *Advances in Neural Information Processing Systems*, 2020.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Birkhäuser, Basel, 2nd edition, 2008.

Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified Q-ensemble. In *Advances in Neural Information Processing Systems*, 2021.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, 2017.

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations*, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223. PMLR, 2017.

Kavosh Asadi, Evan Cater, Dipendra Misra, and Michael L Littman. Equivalence between Wasserstein and value-aware loss for model-based reinforcement learning. *arXiv preprint arXiv:1806.01265*, 2018.

Arip Asadulaev, Rostislav Korst, Aleksandr Korotin, Vage Egiazarian, Andrey Filchenkov, and Evgeny Burnaev. Rethinking optimal transport in offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2024.

Ali Baheri. Risk-aware reinforcement learning through optimal transport theory. *arXiv preprint arXiv:2309.06239*, 2023a.

Ali Baheri. Understanding reward ambiguity through optimal transport theory in inverse reinforcement learning. *arXiv preprint arXiv:2310.12055*, 2023b.

Ali Baheri. Wasserstein-barycenter consensus for cooperative multi-agent reinforcement learning. *arXiv preprint arXiv:2506.12497*, 2025.

Ali Baheri and Mykel J Kochenderfer. The synergy between optimal transport theory and multi-agent reinforcement learning. *arXiv preprint arXiv:2401.10949*, 2024.

Ali Baheri, Zahra Sharooei, and Chirayu Salgarkar. Wave: Wasserstein adaptive value estimation for actor-critic reinforcement learning. In *7th Annual Conference on Learning for Dynamics and Control*, volume 283, pp. 1–12. PMLR, 2025.

Alex Beeson and Giovanni Montana. Balancing policy constraint and ensemble size in uncertainty-based offline reinforcement learning. *Machine Learning*, 113(1):443–488, 2024.

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

Jeremiah Birrell, Paul Dupuis, Markos A Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet. (f, Γ)-divergences: Interpolating between f-divergences and integral probability metrics. *Journal of Machine Learning Research*, 23(39):1–70, 2022.

Michael Bloem and Nicholas Bambos. Infinite time horizon maximum causal entropy inverse reinforcement learning. In *53rd IEEE Conference on Decision and Control*, pp. 4911–4916. IEEE, 2014.

Maksim Bobrin, Nazar Buzun, Dmitrii Krylov, and Dmitry V. Dylov. Align your intents: Offline imitation learning via optimal transport. *arXiv preprint arXiv:2402.13037*, 2024.

Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using Lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, pp. 1–12. ACM, 2011.

David Brandfonbrener, Will Whitney, Rajesh Ranganath, and Joan Bruna. Offline RL without off-policy evaluation. In *Advances in Neural Information Processing Systems*, 2021.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

Alexander Bukharin, Yan Li, Yue Yu, Qingru Zhang, Zhehui Chen, Simiao Zuo, Chao Zhang, Songan Zhang, and Tuo Zhao. Robust multi-agent reinforcement learning via adversarial regularization: Theoretical foundation and stable algorithms. In *Advances in Neural Information Processing Systems*, 2023.

Sergio Calo, Anders Jonsson, Gergely Neu, Ludovic Schwartz, and Javier Segovia-Aguas. Bisimulation metrics are optimal transport distances, and can be computed efficiently. In *Advances in Neural Information Processing Systems*, 2024.

Wei-Di Chang, Scott Fujimoto, David Meger, and Gregory Dudek. Imitation learning from observation through optimal transport. *arXiv preprint arXiv:2310.01632*, 2023.

Laetitia Chapel, Romain Tavenard, and Samuel Vaiter. Differentiable generalized sliced Wasserstein plans. In *Advances in Neural Information Processing Systems*, 2025.

Liqun Chen, Ke Bai, Chenyang Tao, Yizhe Zhang, Guoyin Wang, Wenlin Wang, Ricardo Henao, and Lawrence Carin. Sequence generation with optimal-transport-enhanced reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7512–7520, 2020.

Yanshuo Chen, Zhengmian Hu, Wei Chen, and Heng Huang. Fast and scalable Wasserstein-1 neural optimal transport solver for single-cell perturbation prediction. *Bioinformatics*, 41:i513–i522, 2025.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

Julien Grand Clement and Christian Kroer. First-order methods for Wasserstein distributionally robust MDP. In *International Conference on Machine Learning*, pp. 2010–2019. PMLR, 2021.

Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *International conference on machine learning*, pp. 1282–1289. PMLR, 2019.

Dane Corneil, Wulfram Gerstner, and Johanni Brea. Efficient model-based deep reinforcement learning with variational state tabulation. In *International Conference on Machine Learning*, pp. 1049–1058. PMLR, 2018.

Jingjing Cui, Yuanwei Liu, and Arumugam Nallanathan. Multi-agent reinforcement learning-based resource allocation for UAV networks. *IEEE Transactions on Wireless Communications*, 19(2):729–743, 2019.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013.

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, pp. 2892–2901, 2018b.

Robert Dadashi, Léonard Hussenot, Matthieu Geist, and Olivier Pietquin. Primal Wasserstein imitation learning. In *International Conference on Learning Representations*, 2021a.

Robert Dadashi, Shideh Rezaeifar, Nino Vieillard, Léonard Hussenot, Olivier Pietquin, and Matthieu Geist. Offline reinforcement learning with pseudometric learning. In *International Conference on Machine Learning*, pp. 2307–2318. PMLR, 2021b.

Fabio De Sousa Ribeiro, Ainkaran Santhirasekaram, and Ben Glocker. Counterfactual identifiability via dynamic optimal transport. In *Advances in Neural Information Processing Systems*, 2025.

Florent Delgrange, Ann Nowe, and Guillermo A Pérez. Wasserstein auto-encoded MDPs: Formal verification of efficiently distilled RL policies with many-sided guarantees. In *International Conference on Learning Representations*, 2023a.

Florent Delgrange, Mathieu Reymond, Ann Nowé, and Guillermo A. Pérez. WAE-PCN: Wasserstein-autoencoded Pareto conditioned networks. In *Adaptive and Learning Agents Workshop (ALA) at AAMAS*, 2023b.

Christopher Diehl, Timo Sievernich, Martin Krüger, Frank Hoffmann, and Torsten Bertram. UMBRELLA: Uncertainty-aware model-based offline reinforcement learning leveraging planning. In *NeurIPS 2021 Machine Learning for Autonomous Driving Workshop*, 2021.

Zixuan Dong, Yumi Omori, and Keith Ross. Is optimal transport necessary for inverse reinforcement learning? *arXiv preprint arXiv:2506.06793*, 2025.

Ishan Durugkar, Steven Hansen, Stephen Spencer, and Volodymyr Mnih. Wasserstein distance maximizing intrinsic control. *arXiv preprint arXiv:2110.15331*, 2021.

Richard L Dykstra. An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78(384):837–842, 1983.

Amir-massoud Farahmand, Andre Barreto, and Daniel Nikovski. Value-aware loss function for model-based reinforcement learning. In *Artificial Intelligence and Statistics*, pp. 1486–1494. PMLR, 2017.

Kilian Fatras, Younes Zine, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Learning with minibatch Wasserstein: Asymptotic and gradient properties. In *International Conference on Artificial Intelligence and Statistics*, volume 108. PMLR, 2020.

Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *International Conference on Machine Learning*, pp. 3186–3197. PMLR, 2021.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *UAI*, volume 4, pp. 162–169, 2004.

Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

Arnaud Fickinger, Samuel Cohen, Stuart Russell, and Brandon Amos. Cross-domain imitation learning via optimal transport. In *International Conference on Learning Representations*, 2022.

Rémi Flamary, Cédric Vincent-Cuaz, Nicolas Courty, Alexandre Gramfort, Oleksii Kachaiev, Huy Quang Tran, Laurène David, Clément Bonet, Nathan Cassereau, Théo Gnassounou, Eloi Tanguy, Julie Delon, Antoine Collas, Sonia Mazelet, Laetitia Chapel, Tanguy Kerdoncuff, Xizheng Yu, Matthew Feickert, Paul Krzakala, Tianlin Liu, and Eduardo Fernandes Montesuma. Pot python optimal transport (version 0.9.5), 2024. URL `https://github.com/PythonOT/POT`.

Vincent François-Lavet, Yoshua Bengio, Doina Precup, and Joelle Pineau. Combined reinforcement learning via abstract representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3582–3589, 2019.

Yuwei Fu, Haichao Zhang, Di Wu, Wei Xu, and Benoit Boulet. Robot policy learning with temporal optimal transport reward. In *Advances in Neural Information Processing Systems*, 2024.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062. PMLR, 2019.

Ted Fujimoto, Joshua Suetterlein, Samrat Chatterjee, and Auroop Ganguly. Assessing the impact of distribution shift on reinforcement learning performance. *arXiv preprint arXiv:2402.03590*, 2024.

Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

Zhaolin Gao, Jonathan Chang, Wenhao Zhan, Owen Oertell, Gokul Swamy, Kianté Brantley, Thorsten Joachims, Drew Bagnell, Jason D Lee, and Wen Sun. REBEL: Reinforcement learning via regressing relative rewards. In *Advances in Neural Information Processing Systems*, 2024.

Milena Gazdieva, Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Extremal domain translation with neural optimal transport. In *Advances in Neural Information Processing Systems*, 2023.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617. PMLR, 2018.

Jonathan Geuter, Gregor Kornhardt, Ingimar Tomasson, and Vaios Laschos. Universal neural optimal transport. In *International Conference on Machine Learning*, volume 267, pp. 19196–19232. PMLR, 2025.

Dibya Ghosh, Chethan Anand Bhateja, and Sergey Levine. Reinforcement learning from passive data via latent intentions. In *International Conference on Machine Learning*, volume 202, pp. 11321–11339. PMLR, 2023.

Arash Givchi, Pei Wang, Junqi Wang, and Patrick Shafto. Distributionally-constrained policy optimization via unbalanced optimal transport. *arXiv preprint arXiv:2102.07889*, 2021.

Yong Liang Goh, Wee Sun Lee, Xavier Bresson, Thomas Laurent, and Nicholas Lim. Combining reinforcement learning and optimal transport for the traveling salesman problem. *arXiv preprint arXiv:2203.00903*, 2022.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

Xiang Gu, Yucheng Yang, Wei Zeng, Jian Sun, and Zongben Xu. Keypoint-guided optimal transport with applications in heterogeneous domain adaptation. In *Advances in Neural Information Processing Systems*, 2022.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, 2017.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering Atari with discrete world models. In *International Conference on Learning Representations*, 2021.

Siddhant Haldar, Vaibhav Mathur, Denis Yarats, and Lerrel Pinto. Watch and match: Supercharging imitation with regularized optimal transport. In *Conference on Robot Learning*, pp. 32–43. PMLR, 2023.

Shuncheng He, Yuhang Jiang, Hongchang Zhang, Jianzhun Shao, and Xiangyang Ji. Wasserstein unsupervised reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6884–6892, 2022.

Charles A Hepburn and Giovanni Montana. Model-based trajectory stitching for improved behavioural cloning and its applications. *Machine Learning*, 113(2):647–674, 2024.

Charles A Hepburn, Yue Jin, and Giovanni Montana. State-constrained offline reinforcement learning. *Transactions on Machine Learning Research*, 2025.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, 2016.

Linfang Hou, Liang Pang, Xin Hong, Yanyan Lan, Zhiming Ma, and Dawei Yin. Robust reinforcement learning with Wasserstein constraint. *arXiv preprint arXiv:2006.00945*, 2020.

Peide Huang, Mengdi Xu, Jiacheng Zhu, Laixi Shi, Fei Fang, and Ding Zhao. Curriculum reinforcement learning using optimal transport via gradual domain adaptation. In *Advances in Neural Information Processing Systems*, 2022.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, 2019.

Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, volume 162. PMLR, 2022.

Nan Jiang. A note on loss functions and error compounding in model-based reinforcement learning. *arXiv preprint arXiv:2404.09946*, 2024.

Aaron Kandel and Scott J. Moura. Safe Wasserstein constrained deep Q-learning. *arXiv preprint arXiv:2002.03016*, 2020.

Leonid Kantorovitch. On the translocation of masses. *Management science*, 5(1):1–4, 1958.

Mete Kemertas and Allan Jepson. Approximate policy iteration with bisimulation metrics. *Transactions on Machine Learning Research*, 2022.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. MOReL: Model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

Pascal Klink, Haoyi Yang, Carlo D'Eramo, Jan Peters, and Joni Pajarinen. Curriculum reinforcement learning via constrained optimal transport. In *International Conference on Machine Learning*, pp. 11341–11358. PMLR, 2022.

Pascal Klink, Carlo D'Eramo, Jan Peters, and Joni Pajarinen. On the benefit of optimal transport for curriculum reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46 (11):7191–7204, 2024.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced Wasserstein distances. In *Advances in Neural Information Processing Systems*, 2019.

Lingkai Kong, Haichuan Wang, Tonghan Wang, Guojun Xiong, and Milind Tambe. Composite flow matching for reinforcement learning with shifted-dynamics data. In *Advances in Neural Information Processing Systems*, 2025.

Cevahir Köprülü, Thiago D. Simão, Nils Jansen, and Ufuk Topcu. Safety-prioritizing curricula for constrained reinforcement learning. In *International Conference on Learning Representations*, 2025.

Arash Bahari Kordabad, Rafael Wisniewski, and Sebastien Gros. Safe reinforcement learning using Wasserstein distributionally robust MPC and chance constraint. *IEEE Access*, 10:130058–130067, 2022.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. In *International Conference on Learning Representations*, 2023.

Ilya Kostrikov, Rob Fergus, Jonathan Tompson, and Ofir Nachum. Offline reinforcement learning with Fisher divergence critic regularization. In *International Conference on Machine Learning*, pp. 5774–5783. PMLR, 2021.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations*, 2022.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy Q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, 2019.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning: State-of-the-art*, pp. 45–73. Springer, 2012.

An Thai Le, Kay Hansel, Jan Peters, and Georgia Chalvatzaki. Hierarchical policy blending as optimal transport. In *Learning for Dynamics and Control Conference*, pp. 797–812. PMLR, 2023.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Haorui Li, Jiaqi Liang, Linjing Li, and Daniel Zeng. Wasserstein diversity-enriched regularizer for hierarchical reinforcement learning. In *International Conference on Neural Information Processing*, pp. 564–577. Springer, 2023.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1192–1202, 2016.

Meng Li, Guangda Huzhang, Haibo Zhang, Xiting Wang, and Anxiang Zeng. Optimal transport-based token weighting scheme for enhanced preference optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21311–21334. Association for Computational Linguistics, 2025.

Wuchen Li and Guido Montúfar. Natural gradient via optimal transport. *Information Geometry*, 1(2): 181–214, 2018.

Amarildo Likmeta, Matteo Sacco, Alberto Maria Metelli, and Marcello Restelli. Wasserstein actor-critic: directed exploration via optimism for continuous-actions control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8782–8790, 2023.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Hanbing Liu, Huaze Tang, Yanru Wu, Yang Li, and Xiao-Ping Zhang. Reinforced domain selection for continuous domain adaptation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Runze Liu, Yali Du, Fengshuo Bai, Jiafei Lyu, and Xiu Li. PEARL: Zero-shot cross-task preference alignment and robust reward learning for robotic manipulation. In *International Conference on Machine Learning*, 2024.

Yicheng Luo, Zhengyao Jiang, Samuel Cohen, Edward Grefenstette, and Marc Peter Deisenroth. Optimal transport for offline imitation learning. In *International Conference on Learning Representations*, 2023.

Nguyen Cong Luong, Dinh Thai Hoang, Shimin Gong, Dusit Niyato, Ping Wang, Ying-Chang Liang, and Dong In Kim. Applications of deep reinforcement learning in communications and networking: A survey. *IEEE Communications Surveys & Tutorials*, 21(4):3133–3174, 2019.

Jiafei Lyu, Mengbei Yan, Zhongjian Qiao, Runze Liu, Xiaoteng Ma, Deheng Ye, Jing-Wen Yang, Zongqing Lu, and Xiu Li. Cross-domain offline policy adaptation with optimal transport and dataset constraint. In *International Conference on Learning Representations*, 2025.

Guillaume Mahey, Laetitia Chapel, Gilles Gasso, Clément Bonet, and Nicolas Courty. Fast optimal transport through sliced generalized Wasserstein geodesics. In *Advances in Neural Information Processing Systems*, 2023.

Ashok Vardhan Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason D. Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, volume 119, pp. 6672–6681. PMLR, 2020.

Igor Melnyk, Youssef Mroueh, Brian Belgodere, Mattia Rigotti, Apoorva Nitsure, Mikhail Yurochkin, Kristjan Greenewald, Jiri Navratil, and Jerret Ross. Distributional preference alignment of LLMs via optimal transport. In *Advances in Neural Information Processing Systems*, 2024.

Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011.

Alberto Maria Metelli, Amarildo Likmeta, and Marcello Restelli. Propagating uncertainty in reinforcement learning via Wasserstein barycenters. In *Advances in Neural Information Processing Systems*, 2019.

David Millard and Ali Baheri. Can optimal transport improve federated inverse reinforcement learning? *arXiv preprint arXiv:2601.00309*, 2026.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pp. 666–704, 1781.

Ted Moskovitz, Michael Arbel, Ferenc Huszar, and Arthur Gretton. Efficient Wasserstein natural gradients for reinforcement learning. In *International Conference on Learning Representations*, 2021.

Vrettos Moulos. Bicausal optimal transport for markov chains via dynamic programming. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pp. 1688–1693. IEEE, 2021.

Byeonghu Na, Hyungho Na, Yeongmin Kim, Suhyeon Jo, HeeSun Bae, Mina Kang, and Il-Chul Moon. Semantic-aware Wasserstein policy regularization for large language model alignment. In *International Conference on Learning Representations*, 2026.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, 2019.

Reabetswe M Nkhumise, Debabrota Basu, Tony J Prescott, and Aditya Gilra. Studying exploration in RL: An optimal transport analysis of occupancy measure trajectories. *Transactions on Machine Learning Research*, 2025.

Motoki Omura, Yusuke Mukuta, Kazuki Ota, Takayuki Osa, and Tatsuya Harada. Offline reinforcement learning with Wasserstein regularization via optimal transport maps. *arXiv preprint arXiv:2507.10843*, 2025.

Aldo Pacchiano, Jack Parker-Holder, Yunhao Tang, Krzysztof Choromanski, Anna Choromanska, and Michael Jordan. Learning to score behaviors for guided policy optimization. In *International Conference on Machine Learning*, pp. 7445–7454. PMLR, 2020.

Zecheng Pan, Zhikang Chen, Ding Li, Min Zhang, Sen Cui, Hongshuo Jin, Luqi Tao, Yi Yang, Deheng Ye, Yu Zhang, Tingting Zhu, and Tianling Ren. Merging without forgetting: Continual fusion of task-specific models via optimal transport. *arXiv preprint arXiv:2511.19561*, 2025.

Georgios Papagiannis and Yunpeng Li. Imitation learning with Sinkhorn distances. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 116–131. Springer, 2022.

Alexandros Papangelis, Yi-Chia Wang, Piero Molino, and Gokhan Tur. Collaborative multi-agent dialogue model training via reinforcement learning. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 92–102, 2019.

Jack Parker-Holder, Aldo Pacchiano, Krzysztof Choromanski, and Stephen Roberts. Effective diversity in population based reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

Deepak Pathak, Parsa Mahmoudieh, Guanghao Luo, Pulkit Agrawal, Dian Chen, Yide Shentu, Evan Shelhamer, Jitendra Malik, Alexei A Efros, and Trevor Darrell. Zero-shot visual imitation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2050–2053, 2018.

Luiz Manella Pereira and M. Hadi Amini. Heterogeneous federated reinforcement learning using Wasserstein barycenters. *arXiv preprint arXiv:2506.15825*, 2025.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pp. 2664–2672. PMLR, 2016.

David Pfau, Ian Davies, Diana Borsa, Joao GM Araujo, Brendan Tracey, and Hado van Hasselt. Wasserstein policy optimization. In *International Conference on Machine Learning*, 2025.

Zhengling Qi, Chenjia Bai, Zhaoran Wang, and Lan Wang. Distributional off-policy evaluation in reinforcement learning. *Journal of the American Statistical Association*, 120(551):1517–1530, 2025.

James Queeney, Erhan Can Ozcan, Ioannis Ch Paschalidis, and Christos G Cassandras. Optimal transport perturbations for safe reinforcement learning with robustness guarantees. *Transactions on Machine Learning Research*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

Zhizhou Ren, Kefan Dong, Yuan Zhou, Qiang Liu, and Jian Peng. Exploration via hindsight goal generation. In *Advances in Neural Information Processing Systems*, 2019.

Pierre H Richemond and Brendan Maginnis. On Wasserstein reinforcement learning and the Fokker-Planck equation. *arXiv preprint arXiv:1712.07185*, 2017.

Marc Rigter, Bruno Lacerda, and Nick Hawes. Rambo-RL: Robust adversarial model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

Willem Röpke, Raphaël Avalos, Roxana Rădulescu, Ann Nowe, Diederik M Roijers, and Florent Delgrange. Integrating RL and planning through optimal transport world models. In *The Seventeenth Workshop on Adaptive and Learning Agents*, 2025.

Mark Rowland, Jiri Hron, Yunhao Tang, Krzysztof Choromanski, Tamas Sarlos, and Adrian Weller. Orthogonal estimation of Wasserstein distances. In *International Conference on Artificial Intelligence and Statistics*, pp. 186–195. PMLR, 2019.

Josh Roy and George D Konidaris. Visual transfer for reinforcement learning via Wasserstein domain confusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9454–9462, 2021.

Srinjoy Roy, Subhajit Saha, and Swagatam Das. From Wasserstein to maximum mean discrepancy barycenters: A novel framework for uncertainty propagation in model-free reinforcement learning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2025.

Thomas Rupf, Marco Bagatella, Nico Gürtler, Jonas Frey, and Georg Martius. Zero-shot offline imitation learning via optimal transport. *arXiv preprint arXiv:2410.08751*, 2024.

Sharan Sahu and Martin T Wells. DRO-REBEL: Distributionally robust relative-reward regression for fast and efficient LLM alignment. *arXiv preprint arXiv:2509.19104*, 2025.

Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving GANs using optimal transport. *arXiv preprint arXiv:1803.05573*, 2018.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Birkhäuser, Cham, 2015.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pp. 1889–1897. PMLR, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Ilana Sebag, Samuel Cohen, and Marc Peter Deisenroth. On combining expert demonstrations in imitation learning via optimal transport. *arXiv preprint arXiv:2307.10810*, 2023.

Zahra Shahrooei and Ali Baheri. Optimal transport-assisted risk-sensitive Q-learning. *arXiv preprint arXiv:2406.11774*, 2024.

Zahra Shahrooei and Ali Baheri. Optimal transport-guided safety in temporal difference reinforcement learning. *arXiv preprint arXiv:2502.16328*, 2025.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 4058–4065, 2018.

Yi Shen, Pan Xu, and Michael M. Zavlanos. Wasserstein distributionally robust policy evaluation and learning for contextual bandits. *Transactions on Machine Learning Research*, 2024.

Noah Y Siegel, Jost Tobias Springenberg, Felix Berkenkamp, Abbas Abdolmaleki, Michael Neunert, Thomas Lampe, Roland Hafner, Nicolas Heess, and Martin Riedmiller. Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. In *International Conference on Learning Representations*, 2020.

Jun Song, Niao He, Lijun Ding, and Chaoyue Zhao. Provably convergent policy optimization via metric-aware trust region methods. *Transactions on Machine Learning Research*, 2023.

Pei-Hao Su, Milica Gasic, Nikola Mrkšić, Lina M Rojas Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. On-line active reward learning for policy optimisation in spoken dialogue systems. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2431–2441, 2016.

Haoyuan Sun and Feng Wu. Less is more: Refining datasets for offline reinforcement learning with reward machines. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 1239–1247, 2023.

Ke Sun, Yingnan Zhao, Wulong Liu, Bei Jiang, and Linglong Kong. Distributional reinforcement learning with regularized Wasserstein loss. In *Advances in Neural Information Processing Systems*, 2024.

Mingyang Sun, Pengxiang Ding, Weinan Zhang, and Donglin Wang. Score-based diffusion policy compatible with reinforcement learning via optimal transport. In *International Conference on Machine Learning*, 2025.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Erik Talvitie. Self-correcting models for model-based reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pp. 2597–2603, 2017.

Antonio Terpin, Nicolas Lanzetti, Batuhan Yardim, Florian Dorfler, and Giorgia Ramponi. Trust region policy optimization with optimal transport discrepancies: Duality and algorithm for continuous actions. In *Advances in Neural Information Processing Systems*, 2022.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2008.

Boyi Wang, Kai Lin, and Guohan Sun. Offline reinforcement learning via optimal transport and improved performance difference theorem. In *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 619–626. IEEE, 2023.

Haozhong Wang, Zhuo Li, Yibo Yang, He Zhao, Hongyuan Zha, and Dandan Guo. Safeguarding LLM fine-tuning via push-pull distributional alignment. *arXiv preprint arXiv:2601.07200*, 2026.

Jianhao Wang, Wenzhe Li, Haozhe Jiang, Guangxiang Zhu, Siyuan Li, and Chongjie Zhang. Offline reinforcement learning with reverse model-based imagination. In *Advances in Neural Information Processing Systems*, 2021.

Kaixin Wang, Bingyi Kang, Jie Shao, and Jiashi Feng. Improving generalization in reinforcement learning with mixture regularization. In *Advances in Neural Information Processing Systems*, 2020.

Keyu Wu, Chao Li, Fengge Wu, and Junsuo Zhao. W-IRL: Inverse reinforcement learning via Wasserstein metric. In *2023 3rd International Conference on Computer, Control and Robotics (ICCCR)*, pp. 291–297. IEEE, 2023.

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

Huang Xiao, Michael Herman, Joerg Wagner, Sebastian Ziesche, Jalal Etesami, and Thai Hong Linh. Wasserstein adversarial imitation learning. *arXiv preprint arXiv:1906.08113*, 2019.

Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. A fast proximal point method for computing exact Wasserstein distance. In *Uncertainty in Artificial Intelligence*, pp. 433–453. PMLR, 2020.

Haoran Xu, Kaiwen Hu, Somayeh Sojoudi, and Amy Zhang. Reinforcement learning via value gradient flow. In *International Conference on Learning Representations*, 2026.

Zaiyan Xu, Sushil Vemuri, Kishan Panaganti, Dileep Kalathil, Rahul Jain, and Deepak Ramachandran. Robust LLM alignment via distributionally robust direct preference optimization. In *Advances in Neural Information Processing Systems*, 2025.

Kai Yan, Alexander Schwing, and Yu-Xiong Wang. Offline imitation from observation via primal Wasserstein state occupancy matching. In *International Conference on Machine Learning*, volume 235, pp. 55814–55838. PMLR, 2024.

Insoon Yang. A convex optimization approach to distributionally robust Markov decision processes with Wasserstein distance. *IEEE Control Systems Letters*, 1(1):164–169, 2017.

Insoon Yang. Wasserstein distributionally robust stochastic control: A data-driven approach. *IEEE Transactions on Automatic Control*, 66(8):3863–3870, 2020.

Siqi Yang, Kai Yan, Alexander G. Schwing, and Yu-Xiong Wang. Latent Wasserstein adversarial imitation learning. *arXiv preprint arXiv:2603.05440*, 2026.

Yucheng Yang, Tianyi Zhou, Qiang He, Lei Han, Mykola Pechenizkiy, and Meng Fang. Task adaptation from skills: Information geometry, disentanglement, and new objectives for unsupervised reinforcement learning. In *International Conference on Learning Representations*, 2025.

Hao Ye, Geoffrey Ye Li, and Biing-Hwang Fred Juang. Deep reinforcement learning based resource allocation for V2V communications. *IEEE Transactions on Vehicular Technology*, 68(4):3163–3173, 2019.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems*, 2020.

Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. COMBO: Conservative offline model-based policy optimization. In *Advances in Neural Information Processing Systems*, 2021.

Xingrui Yu, Zhenglin Wan, David Mark Bossens, Yueming Lyu, Qing Guo, and Ivor W Tsang. Imitation from diverse behaviors: Wasserstein quality diversity imitation learning with single-step archive exploration. *arXiv preprint arXiv:2411.06965*, 2024.

Maryam Zare, Parham M Kebria, and Abbas Khosravi. Leveraging optimal transport for enhanced offline reinforcement learning in surgical robotic environments. *arXiv preprint arXiv:2310.08841*, 2023.

Michael R Zhang, Tom Le Paine, Ofir Nachum, Cosmin Paduraru, George Tucker, Ziyu Wang, and Mohammad Norouzi. Autoregressive dynamics models for offline policy evaluation and optimization. In *International Conference on Learning Representations*, 2021.

Ming Zhang, Yawei Wang, Xiaoteng Ma, Li Xia, Jun Yang, Zhiheng Li, and Xiu Li. Wasserstein distance guided adversarial imitation learning with reward shape exploration. In *2020 IEEE 9th Data Driven Control and Learning Systems Conference (DDCLS)*, pp. 1165–1170. IEEE, 2020.

Ruiyi Zhang, Changyou Chen, Chunyuan Li, and Lawrence Carin. Policy optimization as Wasserstein gradient flows. In *International Conference on Machine Learning*, pp. 5737–5746. PMLR, 2018.

Mo Zhou, Haosheng Zhou, and Ruimeng Hu. Learning mean-field games through mean-field actor-critic flow. *arXiv preprint arXiv:2510.12180*, 2025.

Wenxuan Zhou, Sujay Bajracharya, and David Held. PLAS: Latent action space for offline reinforcement learning. In *Conference on Robot Learning*, pp. 1719–1735. PMLR, 2021.

Yujie Zhu, Charles A Hepburn, Matthew Thorpe, and Giovanni Montana. Uncertainty-based smooth policy regularisation for reinforcement learning with few demonstrations. In *Advances in Neural Information Processing Systems*, 2025.

Appendix: notation and summary tables

Table 2: Notation used throughout the survey. We use lower-case v^π, q^π for true value functions and upper-case V_θ, Q_θ for learned approximations; $R(s,a)$ denotes the reward function and r_t realised rewards.

MDP and RL	
$\mathcal{S}, \mathcal{A}$	State and action spaces.
$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma, p_0)$	Discounted Markov decision process (MDP).
γ	Discount factor.
p_0 (or ν)	Initial-state distribution (default p_0; ν when varying the start distribution).
$P(\cdot \mid s,a)$	Environment transition kernel.
P_π	Trajectory law generated by (p_0, π, P).
$P_0(\cdot \mid s,a),\ P_\phi(\cdot \mid s,a)$	Nominal and learned transition model / world model (parameters ϕ).
$R(s,a),\ r_t$	Reward function and realised reward $r_t := R(s_t, a_t)$.
$\pi,\ \pi_\theta$	Policy (possibly parametrised by θ).
$\pi_\beta,\ \pi_E$	Behaviour/data-collection policy; expert policy.
τ	Trajectory, e.g. $(s_0, a_0, r_0, s_1, \dots)$.
$t,\ T$	Time index; horizon/trajectory length (as used in finite-horizon transports).
G_t	Discounted return from time t.
$J(\pi)$	Expected return objective.
$v^\pi(s),\ q^\pi(s,a)$	State-value and action-value under policy π.
$V_{\theta_v}(s),\ Q_{\theta_q}(s,a)$	Learned approximations (critics) to v^π, q^π.
$A^\pi(s,a)$	Advantage function.
$\rho_\pi(s,a)$	Normalised discounted occupancy measure; $\rho_\pi(s) = \sum_a \rho_\pi(s,a)$.
$\hat{p}(s)$	Empirical state occupancy measure of a single finite trajectory.
$\mathcal{D}$	Offline dataset of transitions (s,a,r,s').
Optimal transport	
$\mathcal{P}(\mathcal{X})$	Probability measures on a space $\mathcal{X}$.
μ, ν	Generic probability measures.
F	Transport map that pushes one probability measure to another.
Γ	Coupling / transport plan on $\mathcal{X} \times \mathcal{Y}$.
$\Pi(\mu,\nu)$	Set of couplings with marginals μ and ν.
$c(x,y),\ d(x,y)$	Ground cost and ground metric.
W_p (e.g. $\mathsf{W}_1, \mathsf{W}_2$)	p-Wasserstein distance (in particular $\mathsf{W}_1, \mathsf{W}_2$).
ε	Entropic regularisation parameter.
W_ε	Entropy regularised Wasserstein distance.
S_ε	Sinkhorn divergence (debiased entropic OT).
$\mathrm{KL}(\cdot\|\cdot)$	Kullback–Leibler divergence.
$u,\ v$	OT dual potentials (Kantorovich dual).
m	Transported mass in partial OT.
λ_1, λ_2	Marginal-penalty weights in unbalanced OT.
η	Wasserstein barycentre.
$d_\mathcal{X}, d_\mathcal{Y};\ \mathrm{GW}$	Intra-space dissimilarities; Gromov–Wasserstein objective.
Imitation / embeddings / distributional RL	
$H(\pi)$	Discounted causal entropy (max-ent IRL).
$\psi;\ \psi^*$	IRL cost regulariser and convex conjugate (as in IRL_ψ).
$\Phi,\ \Phi_\#$	Embedding map and pushforward.
μ_π	Behavioural distribution $\mu_\pi = \Phi_\# P_\pi$.
$\hat{p},\ \hat{p}_\pi,\ \hat{p}_E$	Empirical state marginals of trajectories (learner/expert).
$Z^\pi(s,a)$	Return distribution (distributional RL).

Continued on next page

MDP and RL (continued)	
$\mathcal{T}^\pi$	(Distributional) Bellman operator.
$\mathcal{T}$	Space of trajectories τ.
$\mathcal{E}$	A latent behaviour space.

Table 3: Representative OT uses for imitation learning and reward inference.

Method (area)	**OT role**	**Compared objects**	**OT form**	**Implementation / notes**
IRL and reward ambiguity				
Baheri (2023b) (IRL)	Reward ambiguity analysis	Representations of feasible reward functions	Wasserstein distance	Computes Wasserstein structure on reward representations; no implementation reported.
Adversarial occupancy matching and discriminator-based IL				
WAIL (Xiao et al., 2019), WDAIL (Zhang et al., 2020) (generative adversarial IL)	Discriminator / reward shaping	State–action occupancy measures	W_1 (Euclidean cost)	Kantorovich duality with 1-Lipschitz critic (WGAN-style training).
LWAIL (Yang et al., 2026) (adversarial IL (TD3))	Latent discriminator + reward	Latent state-pair (transition) occupancies (ICVF-pretrained features)	W_1 (Euclidean cost)	Duality-based critic in latent space; pretraining via ICVF as proposed.
SIL (Papagiannis & Li, 2022) (adversarial IL)	Discriminator / reward shaping via primal OT	State–action occupancy measures	Entropy-regularised W_ε (OT-GAN-style cost (Salimans et al., 2018))	Sinkhorn iterations in minibatches; entropic OT objective.
PWIL (Dadashi et al., 2021a) (IL)	Reward-from-transport (non-adversarial)	State–action occupancy measures	Upper bound on W_1	Uses approximate primal coupling (e.g. POT (Flamary et al., 2024)); transport cost converted to rewards.
PW-DICE (Yan et al., 2024) (offline Learning from Observations (LfO))	Weighted BC with potentials solved by OT	State occupancy measures	primal W_1 (a weighted sum of state occupancy with the Euclidean distance between state embeddings as the cost metric)	A smooth convex dual objective
Offline reward labelling and temporal/intent extensions				
OTR (Luo et al., 2023) ((Chang et al., 2023; Haldar et al., 2023; Wang et al., 2023; Zare et al., 2023)) (Offline RL)	Reward labelling	Trajectory state marginals / occupancies	Entropic OT (cosine cost)	Sinkhorn algorithm on state features; labels used for downstream RL.
Dong et al. (2025) (IRL benchmark)	Empirical comparison / baseline analysis	Trajectory state marginals (replicates OTR-style labelling)	Various (reproduces W_1, entropic, cosine)	Argues proximity baselines can match OT-based reward labelling; provides controlled comparisons across cost and solver choices.

Continued on next page

Method (area)	OT role	Compared objects	OT form	Implementation / notes
SMMOTIL (Sebag et al., 2023) (IRL)	Reward labelling	Agent and multiple expert trajectories (state occupancies)	Sliced multi-marginal W_2	Monte Carlo estimation.
OOPS (Chang et al., 2023) (ILfO)	Reward labelling	Trajectory transition occupancies	Entropic W_1 (square root of the Euclidean distance)	Sinkhorn algorithm.
AILOT (Bobrin et al., 2024) (Offline RL)	Reward labelling with intent-aware costs	Intention/state representations	Entropy-regularised W_ε	Sinkhorn algorithm with temporal-dependent cost as proposed.
TemporalOT (Fu et al., 2024) (Offline RL)	Reward labelling with temporal constraints	Visual embeddings with temporal mask/context cost	Masked entropic OT	Sinkhorn algorithm with masked coupling and smoothed costs.
Pretraining and cross-domain pipelines				
OTDF (Lyu et al., 2025) (cross-domain offline RL)	Data filtering / reweighting	State–action–next-state triples	Entropic W_ε / Sinkhorn	Sinkhorn-based scoring used to filter or reweight data (OTR-style).
GWIL (Fickinger et al., 2022) (cross-domain IL)	Cross-domain reward / matching	State–action occupancy measures via relations	GW distance	GW solvers with entropic regularisation (Peyré et al., 2016), building on regularised standard OT (Ferradans et al., 2014).

Table 4: Representative OT methods for online policy optimisation and closely related objectives.

Method (setting)	OT role	Compared objects	OT form	Implementation / notes
Policy update constraints and regularisers				
WPO, SPO (Song et al., 2023) (on-policy; discrete)	Trust-region constraint	Per-state action distributions $\pi'(\cdot \mid s)$ vs. $\pi(\cdot \mid s)$	W_1 / Sinkhorn	Per-state Lagrangian dual; closed-form distribution update under task-dependent action costs (e.g. 0–1, L_1).
OT-TRPO (Terpin et al., 2022) (TRPO-style)	Trust-region constraint	Action distributions (e.g. Gaussians)	W_2	Closed-form W_2 for Gaussians under quadratic costs; dual optimisation over the multiplier.
BGPG, BGES (Pacchiano et al., 2020) (PG / ES)	Behaviour regulariser	Embedding distributions μ_{π_θ} vs. μ_{ref}	Entropy-regularised W_ε	Dual OT test functions on embedding space; OT signal used as shaping/weights in PG or ES updates.
WPR (Na et al., 2026) (RLHF)	Policy regularisation	learned policy π_θ vs. reference policy π_{ref}	Entropy-regularised W_ε	Dual formulation with Sinkhorn-Knopp iterations yields a tractable semantic regulariser compatible with PPO-style updates.
Wasserstein geometry: gradient flows and natural gradients				
WGF (Richemond & Maginnis, 2017) (bandit / 1-step setting)	Gradient flow	Action distributions (bandit / 1-step setting)	W_2	JKO/proximal viewpoint; policy updates via Wasserstein measure evolution.
WGF (Zhang et al., 2018) (trajectory-measure view)	Gradient flow	Policy-induced measures (e.g. trajectory laws)	W_2	JKO/proximal viewpoint; policy updates via Wasserstein measure evolution.

Continued on next page

Method (setting)	OT role	Compared objects	OT form	Implementation / notes
WNG (WNPG / WNES) (Moskovitz et al., 2021) (parametric policies)	Natural gradient	Policy-induced measures / behaviours	W_2	Wasserstein information geometry; kernel/low-rank approximations yield practical natural-gradient steps (policy-gradient and evolution-strategies variants).
Wasserstein policy optimisation (Pfau et al., 2025) (actor–critic)	Gradient-flow update	Policy distributions	W_2	Discretises a Wasserstein gradient flow and projects to a parametric policy class; the implemented algorithm keeps KL regularisation in practice.
Distributional RL				
QR-DQN (Dabney et al., 2018b) (distributional RL)	Bellman updates	Return distributions $Z^\pi(s,a)$ (action–return)	W_1	Quantile-based approximation via fixed quantiles / quantile regression.
SinkhornDRL (Sun et al., 2024) (distributional RL)	Return-distribution update	Particle/empirical return distributions	Sinkhorn divergence	Sinkhorn iterations on particle sets (debiased entropic OT).
Value learning and OT-regularised critics				
WQL (Metelli et al., 2019) (value learning)	Wasserstein TD	Q-posteriors vs. targets	W_2 barycentres	Closed-form barycentre updates under Gaussian/particle approximations; uncertainty propagation via Wasserstein TD.
WAC (Likmeta et al., 2023) (SAC)	Critic update	Gaussian Q-posterior vs. target	W_2	Closed-form W_2 for Gaussian families; actor exploits posterior structure.
WAVE (Baheri et al., 2025) (actor–critic)	Stabilise critic drift	Critic-output distributions across iterations	Entropic OT	Sinkhorn OT penalty between successive critic outputs; adaptive weighting.

Table 5: Representative OT methods for model-based RL and Wasserstein world models.

Method (setting)	OT role	Compared objects	OT form	Implementation / notes
Value-awareness				
VAML (Farahmand et al., 2017; Asadi et al., 2018) (model-based)	Value-aware transition model loss	$P(\cdot \mid s,a)$ vs. $P_\phi(\cdot \mid s,a)$	W_1	Equivalence between Lipschitz test-function losses and a W_1 discrepancy on next-state distributions.
Link to bisimulation metric				
(Calo et al., 2024) (dynamic programming)	Efficient OT distances for Markov processes	Stationary Markov processes (occupancy of state pairs)	Discounted W_1	Sinkhorn value iteration; Markov-structured couplings improve computation and convergence.
Latent MDP training				
WAE-MDP (Delgrange et al., 2023a) (latent world model)	OT-based model learning	Trace/trajectory distributions (encoded/decoded)	W_1-type	WAE/WGAN-style objective; dual-critic realisation with gradient penalty in practice.
OT-MDP (Röpke et al., 2025) (latent world model)	Regularised OT objective	Real vs. latent trajectories / transitions	OT-based	Approximates via total variation; discretised latent states with straight-through estimators (as proposed).

Table 6: Representative OT uses for offline policy optimisation.

Method (area)	**OT role**	**Compared objects**	**OT form**	**Implementation / notes**
BRAC (Wu et al., 2019) (offline RL)	Policy regularisation / value penalty	Action distributions $\pi(\cdot \mid s)$ vs. $\pi_\beta(\cdot \mid s)$	W_1	Dual form with a discriminator (WGAN-style critic).
Q-DOT (Omura et al., 2025) (offline RL)	Policy regularisation via transport map	Action distributions (state-conditional)	W_2	ICNN parameterisation of a convex potential / transport map.
VGF (Xu et al., 2026) (offline RL)	Gradient flow	Action distributions (state-conditional)	W_2	JKO scheme.
PPL (Asadulaev et al., 2024) (offline RL)	Partial behaviour cloning / selective matching	Action distributions under partial OT constraint	Partial OT	Maximin formulation with a tunable partiality parameter.
Givchi et al. (2021) (RL (policy optimisation))	Unbalanced OT formulation	Projected state and action marginals of occupancies	Unbalanced OT	Multiple convex constraints handled via Dykstra's algorithm; Bregman (KL) divergences used in practice.
Fatras et al. (2021) (Domain adaptation)	Domain-invariant representation learning	Source vs. target distributions	Unbalanced minibatch OT	POT and Geomloss packages

Table 7: Representative OT uses for control shift and transfer.

Method (area)	**OT role**	**Compared objects**	**OT form**	**Implementation / notes**
Distributionally robust control				
WR^{2}L (Abdullah et al., 2019) (robust RL)	Minimax constraint	Transition kernels $P(\cdot \mid s, a)$ within a Wasserstein ball	W_2	Approximates W_2 via a Hessian-of-expectation surrogate; zero-order optimisation as proposed.
DRMDP (Kordabad et al., 2022; Yang, 2020; Hou et al., 2020; Queeney et al., 2024; Yang, 2017) (robust control / robust RL)	Ambiguity set (Wasserstein ball)	Disturbance or transition distributions	OT-cost uncertainty set (generalises W_p)	Kantorovich duality yields tractable robust Bellman operators / reformulations.
FOM for DRMDP (Clement & Kroer, 2021) (robust control / robust RL)	Ambiguity set (Wasserstein ball)	Disturbance or transition distributions	W_2	First-order methods.
Safety				
(Baheri, 2023a; Shahrooei & Baheri, 2024) (safe RL)	Visitation shaping / regularisation	Risk distribution vs. state visitation distribution	Wasserstein with squared Euclidean cost	OT-derived costs used as penalties in value-based updates (as proposed).
Shahrooei & Baheri (2025) (safe RL (SARSA))	Action uncertainty scoring	Action distributions induced by current vs. target Q estimates	Entropic OT	Sinkhorn iterations compute uncertainty scores via an OT map; used to define safer behavioural policy.
RLHF / preference alignment				
Xu et al. (2025), Sahu & Wells (2025) (RLHF)	Wasserstein DRO on preferences	Preference distributions (nominal vs. shifted)	W_p ball	Robust DPO/REBEL with Wasserstein ambiguity sets over preference distributions.

Continued on next page

Method (area)	OT role	Compared objects	OT form	Implementation / notes
OTPO (Li et al., 2025) (RLHF)	Token-level saliency weighting	Token representations within preference pairs	Unbalanced entropic OT	OT couplings assign importance weights to tokens; biases optimisation toward preference-driving content.
SOT (Wang et al., 2026) (safety-aware SFT, not RLHF)	Push–pull distributional alignment	Model output vs. desired/undesired reference distributions	OT-based weighting	Dual-reference OT: pushes toward desired, pulls from undesired distribution during fine-tuning.
Transfer				
ZILOT (Rupf et al., 2024) (zero-shot offline IL)	Goal/intent alignment for transfer	State and goal occupancy measures	W_1 (goal-conditioned cost)	Sinkhorn algorithm with a goal-conditioned value function used as ground cost.
PEARL (Liu et al., 2024) (preference-based RL)	Preference transfer	Trajectory sets via preference relations	GW	Sinkhorn-style solver for GW alignment; transfers preferences by matching intra-trajectory structure.
WDGRL (Shen et al., 2018) (domain adaptation)	Domain-invariant representation learning	Feature distributions (source vs. target)	W_1	Adversarial Wasserstein alignment (WGAN-style duality).
WAPPO (Roy & Konidaris, 2021) (RL under visual shift)	Domain-agnostic representation for control	Feature distributions across domains	W_1	Dual formulation; minimises feature-distribution Wasserstein distance during policy learning.
Liu et al. (2025) (domain selection)	Reward shaping for intermediate-domain choice	Domain-specific feature distributions	Wasserstein distance	Uses Wasserstein-based reward to guide selection of intermediate domains (as proposed).
Curricula				
CURROT (Klink et al., 2022; 2024) (curriculum RL)	Curriculum generation	Task/context distributions	Constrained W_2 distance	Linear assignment between particle sets; shortest-path / performance pseudo-distances as ground costs.
GRADIENT (Huang et al., 2022) (curriculum RL)	Curriculum generation	Context/task distributions	Wasserstein barycentre (W_1)	Sinkhorn algorithm under a π-contextual distance.
HGG (Ren et al., 2019) (goal-conditioned RL)	Goal generation for exploration	Joint distribution of initial state and goal	W_1	Maximum-weight bipartite matching for goal selection under a Wasserstein objective.

Table 8: Representative OT uses in multi-agent RL and other applications.

Method (area)	OT role	Compared objects	OT form	Implementation / notes
Diversity and hierarchical RL				
WURL (He et al., 2022) (unsupervised RL)	Intrinsic reward for diversity	State marginal distributions across policies/skills	W_1	Dual form or projected Wasserstein discrepancy; can amortise rewards from OT plans.
PWSEP (Yang et al., 2025) (unsupervised RL)	Skill separation / discovery	State distributions conditioned on skills	Wasserstein (sliced approximation)	Uses sliced Wasserstein distance for efficient estimation of separability objective.

Continued on next page

Method (area)	**OT role**	**Compared objects**	**OT form**	**Implementation / notes**
WQDIL (Yu et al., 2024) (quality-diversity IL)	Stable reward learning in latent space	Latent distributions of demonstrations vs. policy trajectories	WAE–WGAN	Duality-based adversarial training within WAE latent space (as proposed).
WDER (Li et al., 2023) (hierarchical RL)	Diversity regulariser for subpolicies	Behavioural embeddings / action distributions	Smoothed W_ε with random-feature approximation	Duality with random-feature approximations enabling scalable updates.
HiPBOT (Le et al., 2023) (hierarchical control)	Policy blending (mixture weights)	Expert-policy and agent-policy priors (temperature-weighted)	Unbalanced entropic OT	Sinkhorn-like scaling algorithm (Chizat et al., 2018).
Multi-agent RL				
Baheri & Kochenderfer (2024) (MARL)	Conceptual integration points	Policy-induced distributions over state/action (various)	W_p / Sinkhorn (various)	Overview of potential OT roles for coordination, scalability, and robustness (position paper).
Baheri (2025) (cooperative MARL)	Coordination via barycentre regularisation	Joint state–action visitation measures	Sinkhorn divergence barycentre	Iterative barycentre computation; sliced-Wasserstein approximation used for scalability.
ERNIE (Bukharin et al., 2023) (MARL)	Robustness (mean-field idea)	Distributions over states (mean-field approximation)	Wasserstein (proposed)	Proposes adopting Wasserstein distance in the mean-field constraint; not implemented in practice (as noted by authors).
Other applications				
Goh et al. (2022) (combinatorial RL)	OT-based decoder in RL pipeline	Permutation/assignment structure (e.g. tours)	Sinkhorn-based	Sinkhorn decoder for non-autoregressive generation in a travelling salesman setting.
OTRL (Chen et al., 2020) (sequence generation)	Regularisation for policy updates	Sequence embeddings vs. reference embeddings	W_1 (cosine cost)	IPOT algorithm (Xie et al., 2020).
OTPR (Sun et al., 2025) (diffusion policy (IQL))	Action reweighting	Masked OT between state/action marginals with Q-based cost	L_2-reg. masked OT	OT-derived compatibility weights reweight the diffusion score loss; masked dual formulation.
Nkhumise et al. (2025) (RL evaluation)	Measure exploration / learning efficiency	State–action occupancy measures across policies	W_1 (L_1 or L_2 ground norm)	Uses OT dataset distance (OTDD) (Alvarez-Melis & Fusi, 2020) to quantify policy update effort and effectiveness.
WIC (Durugkar et al., 2021) (Unsupervised RL)	Reward shaping for skill-conditioned policies learning	State visitation distribution vs. start state distribution	W_1 (time-step metric)	Duality with a 1-Lipschitz potential.
Wasserstein DRO Shen et al. (2024) (Contextual bandits)	Ambiguity set for robust policy evaluation/learning	Candidate vs. empirical context distributions	$\mathsf{W}_2/\mathsf{W}_\varepsilon$	Dual problem.