Learning from the Right Mistakes: When Do Low-Performing Data Help Offline Policy Gradients?

Anonymous Authors[1]

Abstract

Imitation learning approaches currently reign supreme for robotic manipulation, as value-based offline reinforcement learning (RL) approaches have not yet proven successful at scaling to state-of-the-art large models. As a result, most robotics training pipelines will filter out subpar data or include them in supervised training objectives. Motivated by recent successes in post-training for large language models, we investigate the viability of policy gradient algorithms with importance sampling as a method of learning from subpar data. We show that with certain dataset compositions common in practical settings, such algorithms can extract useful additional learning signal from low-performing data. We also find that such approaches are not able to extract useful signal from low-performing data within datasets that are formed from the replay buffers of agents trained with RL, a dataset composition that is prevalent in the offline RL literature but rare in the real world. Our results point to the importance of considering the interplay between dataset composition and offline RL algorithm design.

1. Introduction

Robotics capabilities have improved massively in the last year due to advances from scaling up imitation learning (Fang et al., 2026; Team et al., 2025). The primary cost of such methods is that they require high-quality data that is scarcely available, often requiring expensive teleoperation setups, highly skilled operators, and significant time investment. The teleoperation data collection process often generates some amount of subpar data as byproducts, but current state-of-the-art training pipelines will simply filter out low-performing data (Chen et al., 2026) or incorporate it through "RL via Supervised Learning" approaches (Emmons et al., 2022; Intelligence et al., 2026).

Both aforementioned approaches are unsatisfying: discarding data is wasteful, and RL via Supervised Learning methods have trouble operating in stochastic environments (Brandfonbrener et al., 2022). We study incorporating low-performing data with binary rewards, operating within the framework of offline RL (Lange et al., 2012; Levine et al., 2020), as online RL methods are sample-inefficient and costly to instrument in real-world settings (Chen et al., 2025; Li et al., 2026). Specifically, we choose to use policy gradient methods, both due to their stability and recent success when applied to large language models (Devvrit et al., 2026; MiniMax et al., 2025) and because learning accurate value functions can be difficult (Li et al., 2023; Nauman et al., 2024; Kumar et al., 2020).

We find that when the low-performing data share some meaningful structure with the high-performing demonstrations, such policy gradient methods can extract useful additional signal from the low-performing data, even without value functions. Interestingly, we do not see the same benefits when the low-performing data is sourced from from the replay buffer of an RL agent, as is typical of many offline RL benchmarks.

2. Preliminaries

We consider the standard offline RL setting (Levine et al., 2020) in which we have a Markov Decision Process (Puterman, 2014) with states $x \in \mathcal{X}$, actions $a \in \mathcal{A}$, transition dynamics $\mathcal{P} : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$. For simplicity, we assume that each trajectory $\tau = (x_0, a_0, x_1, \ldots, x_T)$ is assigned a binary scalar reward $R(\tau) \in \{1, -1\}$ corresponding to success or failure. Given an offline dataset $\mathcal{D} = \{\tau_i\}_i$, the goal is to learn a policy $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ that maximizes the expected reward: $J(\pi) = \mathbb{E}_{\tau\sim\pi}[R(\tau)]$.

We choose Tapered Off-Policy REINFORCE (TOPR; Le Roux et al. (2025)) as a prototypical policy gradient algorithm that can stably incorporate off-policy data. With

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

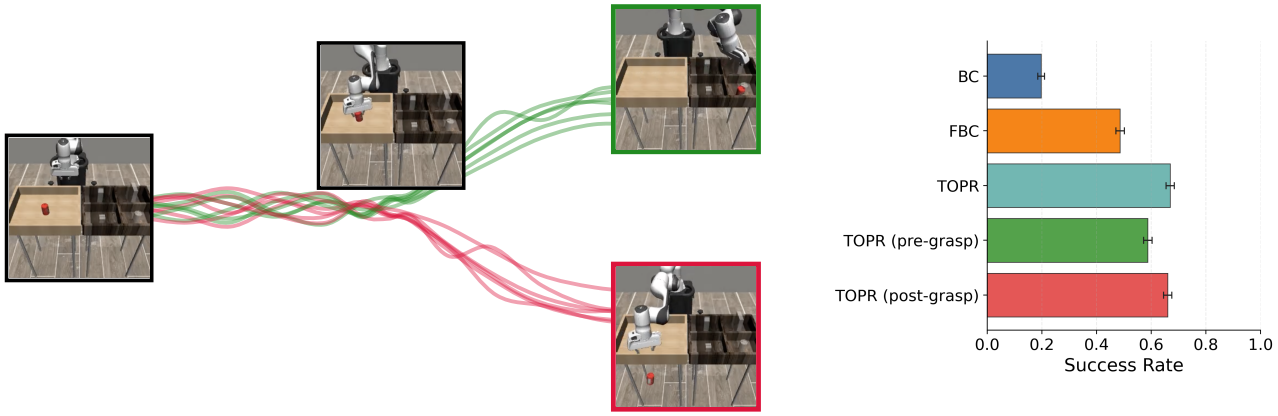

Figure 1. Left: Illustration of the *shared prefix* structure in `can-paired`. All trajectories successfully grasp the can, after which half of the rollouts place the can in the proper bin (**successful** trajectories) while the other half drop the can off of the table (**failed** trajectories). **Right:** TOPR is able to extract useful signal from low-performing trajectories, allowing it to outperform FBC. TOPR using only $\nabla J^-_{\mathcal{M}^{\text{pre}}}$ yields worse performance due to interference, while using $\nabla J^-_{\mathcal{M}^{\text{post}}}$ performs nearly identically to using the full ∇J^-.

binary rewards, TOPR uses the asymmetric policy gradient

$$\nabla J(\pi) = \underbrace{\mathbb{E}_{\tau\sim\mathcal{D}^+}\left[\nabla \log \pi(\tau)\right]}_{\nabla J^+} \underbrace{-\mathbb{E}_{\tau\sim\mathcal{D}^-}\left[[\rho(\tau)]_0^1 \nabla \log \pi(\tau)\right]}_{\nabla J^-}, \quad (1)$$

where $\mathcal{D}^+ = \{\tau \in \mathcal{D} : R(\tau) = 1\}$ is the set of successful trajectories, $\mathcal{D}^-$ is symmetrically defined as the set of failed trajectories, and $[x]_0^1 := \max(0, \min(1, x))$ clips the trajectory-level importance ratios $\rho(\tau)$:

$$\rho(\tau) = \prod_{t=0}^{T-1} \frac{\pi(a_t \mid x_t)}{\pi_{\text{ref}}(a_t \mid x_t)}.$$

This clipping helps to avoid the high variance of the product of many per-timestep ratios (Precup et al., 2000). The reference policy is frozen and initialized at the beginning of training as $\pi_{\text{ref}} \leftarrow \pi$. Le Roux et al. (2025) construct $\mathcal{D}$ by collecting on-policy rollouts from π; in their case, the reference policy π_{ref} also functions as the data-generating policy. In general, these may not be the same, as it may not even be possible to access the data-generating policy beyond the collected trajectories (e.g., if $\mathcal{D}$ is human-generated).

The decomposition of the TOPR objective (1) into ∇J^+ and ∇J^- is illustrative: ∇J^+ corresponds to performing behavior cloning (BC; Pomerleau (1988)) on the successful trajectories and ∇J^- corresponds to truncated importance sampling (Munos et al., 2016; Espeholt et al., 2018) for the failed trajectories. A natural baseline approach is to only train with the term ∇J^+, which we will refer to as *filtered behavior cloning* (FBC; Chen et al. (2021)).

3. The Impact of "Shared Prefix" Data Structure

We seek to understand the conditions under which low-performing offline data can be helpful. To this end, we will highlight that the shared structure between the "positive" and "negative" data is of central importance, and in some cases can lead to harmful interference (McCloskey & Cohen, 1989) effects. To illustrate these phenomena, we study the `can-paired` environment from the `robomimic` task suite (Mandlekar et al., 2021). `can-paired` consists of paired human teleoperation demonstrations collected from the same initial state: for each initial state, one trajectory will successfully complete the task of picking up the can and placing it in the correct receptacle. The other trajectory from that initial state will pick up the can and drop it off of the edge of the table (Fig. 1, left). Importantly, the failed trajectories exhibit proper grasping behavior (i.e., the pre-grasp portions of all trajectories are approximately identically distributed). We will refer to this property as the data possessing a *shared prefix* structure; the failed trajectories begin with behavior resembling that of the successful demonstrations.

The potential trouble with such shared prefix structure is that there is a portion of the trajectories in $\mathcal{D}^+$ and $\mathcal{D}^-$ for which ∇J^+ is trying to maximize the likelihood while ∇J^- is trying to minimize the likelihood of data coming from that same distribution. To be more precise, consider binary masks of the form $\mathcal{M}^\tau \in [0,1]^T$ to identify subsets of timesteps, with $\mathcal{M}_i^{\text{pre}} = \mathbb{1}_{\{i<T_g\}}$ indicating the timesteps before the grasp time T_g and $\mathcal{M}_i^{\text{post}} = \mathbb{1}_{\{i\geq T_g\}}$ indicating the post-grasp timesteps. These masks let us rea-

son about ∇J^- when restricted to specific subsegments:

$$\nabla J^-_{\mathcal{M}} = -\mathbb{E}_{\tau\sim\mathcal{D}^-}\left[[\rho(\tau)]_0^1 \sum_{t=0}^{T-1} \mathcal{M}_t \nabla \log \pi(a_t|x_t)\right],$$

with $\mathcal{M} \in \{\mathcal{M}^{\text{pre}}, \mathcal{M}^{\text{post}}\}$.

In this notation, we can hypothesize that shared prefix structure will lead to gradient interference, manifested in the form of a large negative cosine similarity between the gradient terms: $\cos(\nabla J^+, \nabla J^-) = \frac{\langle \nabla J^+, \nabla J^- \rangle}{\|\nabla J^+\| \|\nabla J^-\|}$. Specifically, we expect $\cos(\nabla J^+, \nabla J^-_{\mathcal{M}^{\text{pre}}})$ to be smaller than $\cos(\nabla J^+, \nabla J^-_{\mathcal{M}^{\text{post}}})$, as ∇J^+ and $\nabla J^-_{\mathcal{M}^{\text{pre}}}$ are opposing objectives on similar data. Interference phenomena are commonplace in the multi-task learning literature, as is the use of gradient cosine similarity as a metric (Yu et al., 2020; Liu et al., 2021).

4. Experiments

To illustrate gradient interference behavior and demonstrate the impact of the shared prefix structure, we conduct the following training-time intervention: train two policies on `can-paired` with TOPR, with one run using $\nabla J^-_{\mathcal{M}^{\text{pre}}}$ to only include timesteps occurring *before* the grasp has been made ("TOPR (pre-grasp)") while the other uses $\nabla J^-_{\mathcal{M}^{\text{post}}}$ for timesteps occurring *after* the grasp ("TOPR (post-grasp)"). Both policies always use the full positive gradient term ∇J^+. In Fig. 2, we see that $\nabla J^-_{\mathcal{M}^{\text{pre}}}$ interferes significantly with ∇J^+. Conversely, $\nabla J^-_{\mathcal{M}^{\text{post}}}$ and ∇J^+ hardly interfere at all in comparison. The interference decreases over the course of training as π moves away from π_{ref} and the importance ratios in ∇J^- decrease.

To expand a bit on this self-correcting mechanism, which may initially seem counterintuitive, recall that the importance ratios $\rho(\tau)$ are computed over the entire trajectory. While the conflicting signal on the shared prefix may prevent the policy from decreasing the likelihood of actions in that shared prefix, there is nothing preventing the policy from decreasing the likelihood of the later actions in the trajectory. Because those timesteps contribute to the trajectory-level ratios $\rho(\tau)$, the weight of ∇J^- is reduced.

Despite the presence of interference, TOPR outperforms FBC on this dataset (Fig. 1, right). Now we address the question of what portion(s) of these failed trajectories provide useful learning signal to account for this difference. Fig. 1 (right) also reports the results for the two masked variants of TOPR analyzed previously. As expected, using $\nabla J^-_{\mathcal{M}^{\text{pre}}}$ is not as useful, and yields performance closer to that of FBC. Further, using $\nabla J^-_{\mathcal{M}^{\text{post}}}$ yields nearly identical performance to baseline TOPR (which applies ∇J^- everywhere). This strongly indicates that it is the portion of the failed trajectories after the grasp that is providing beneficial

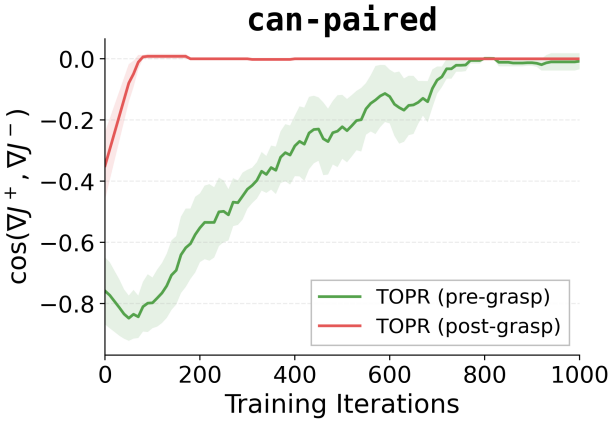

Figure 2. Cosine similarity of ∇J^+ and ∇J^- on `can-paired`. ∇J^+ and $\nabla J^-_{\mathcal{M}^{\text{pre}}}$ point in nearly opposite directions near the beginning of training (**green**), while $\nabla J^-_{\mathcal{M}^{\text{post}}}$ does not interfere much with ∇J^+ (**red**).

learning signal. Since the shared prefix just seems to be a source of potential optimization woes for this dataset, we are motivated to ask the question: are there greater benefits in datasets without shared prefix structure?

To answer this question, we further our analysis with two `robomimic` datasets containing highly unstructured failed trajectories: `can-mg` and `lift-mg`, which are comprised of rollouts collected by a Soft Actor-Critic (Haarnoja et al., 2018) agent during its training process. The goal in `can-mg` is identical to that of `can-paired`, while the goal in `lift-mg` is to grasp and pick up a cube from a table. The distributions of trajectories in these datasets are thus similar to those of an agent's replay buffer, which is a popular dataset composition in the offline RL literature (Fu et al., 2021). These datasets do not possess the same qualitatively obvious shared prefix structure as `can-paired`; most of the failed trajectories consist of the robot arm flailing around, and do not resemble the successful behavior at all. Across both `can-mg` and `lift-mg`, TOPR performs nearly identically to FBC (Fig. 3, top). It is perhaps unsurprising that TOPR does not perform appreciably better than FBC, as the failed trajectories provide little information near the support of the state distribution of a successful policy. Interestingly, training on both of these datasets leads to strong interference dynamics (Fig. 3, bottom) which suggests that there may be additional mechanisms at play contributing to these phenomena.

5. Discussion

We wish to contextualize our findings in the existing offline RL literature. Most prior works in value-based of-

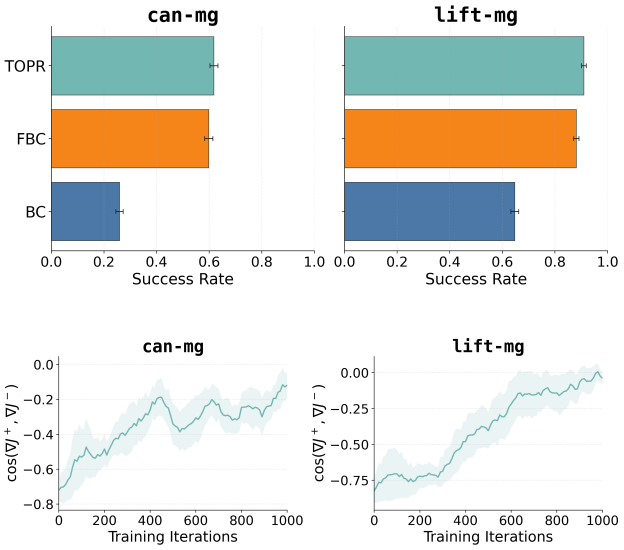

Figure 3. Top: TOPR and FBC perform similarly on datasets resembling RL agent replay buffers. **Bottom:** TOPR training encounters significant gradient interference on both environments despite no overt presence of shared prefix structure in the dataset composition of `can-mg` and `lift-mg`.

fline RL use methods based around the intuitive idea of "stitching together" good segments of low-performing trajectories (Kostrikov et al., 2022). While this is a sensible way to learn from subpar data, it is almost the opposite of what we demonstrate in this paper. To illustrate the difference, consider that value-based methods could recognize that the shared prefixes in `can-paired` consist of good actions, and can use those to improve the policy. In contrast, TOPR mostly finds learning signal not from the good pre-grasp data in failed trajectories, but from the bad post-grasp continuations. This points to a limitation of methods like TOPR that do not have any mechanisms for fine-grained credit assignment: potential benefits from cloning the high-quality prefixes of the failed trajectories are lost.

Our experiment in Section 4 using only $\nabla J^-_{\mathcal{M}^{\text{post}}}$ to reduce gradient interference was presented primarily as a diagnostic experiment, but could potentially be explored as a practical approach. Choosing to limit the mask to only post-grasp timesteps can be thought of as a form of manual credit assignment, which in this case came from prior domain knowledge. This is important to acknowledge, but this does not necessarily render the approach intractable for real-world use; this method could be enabled with coarsely-annotated segments of data containing "mistakes" identified by human labelers (Intelligence et al., 2026) or reward models (Chen et al., 2026). While these two annotation methods have their own drawbacks, they may be easier to scale than learning value functions for credit assignment (Park et al., 2025).

Our findings relating gradient interference and shared prefix structure are intriguing, considering that policy gradient algorithms with importance sampling are very effective for large language model post-training (MiniMax et al., 2025), a domain in which shared prefix structures are known to be abundant (Wang et al., 2025). As large models for robotic control continue to improve in base capabilities, we expect failures to look less like "flailing around" and more like failures in the reasoning chain of a language model, in which one or more mistakes is contained within a long sequence of subtasks. Some prior works explicitly encourage collecting demonstrations with intentional mistakes followed by corrective behavior (Hu et al., 2025); we note that this is just another example of specially structured offline data. We hope that our findings encourage careful consideration of the interaction between the structure of offline data and the algorithms we use for offline learning, and that new patterns of dataset composition may spark the design of new strains of offline RL algorithms suited to those patterns.

References

Brandfonbrener, D., Bietti, A., Buckman, J., Laroche, R., and Bruna, J. When does return-conditioned supervised learning work for offline reinforcement learning? *Advances in Neural Information Processing Systems*, 35:1542–1553, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0a2f65c9d2313b71005e600bd23393fe-Abstract-Conference.html.

Chen, K., Liu, Z., Zhang, T., Guo, Z., Xu, S., Lin, H., Zang, H., Li, X., Zhang, Q., Yu, Z., Fan, G., Huang, T., Wang, Y., and Yu, C. $\pi_{\texttt{rl}}$: Online RL Fine-tuning for Flow-based Vision-Language-Action Models, October 2025. URL http://arxiv.org/abs/2510.25889. arXiv:2510.25889 [cs].

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. Decision Transformer: Reinforcement Learning via Sequence Modeling. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15084–15097. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/7f489f642a0ddb10272b5c31057f0663-Abstract.html.

Chen, Q., Yu, J., Schwager, M., Abbeel, P., Shentu, F., and Wu, P. SARM: Stage-aware reward modeling for long horizon robot manipulation. In *The Fourteenth*

International Conference on Learning Representations, 2026. URL https://openreview.net/forum?id=aemqAxScl9.

Devvrit, F., Madaan, L., Tiwari, R., Bansal, R., Duvvuri, S. S., Zaheer, M., Dhillon, I. S., Brandfonbrener, D., and Agarwal, R. The art of scaling reinforcement learning compute for LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=FMjeC9Msws.

Emmons, S., Eysenbach, B., Kostrikov, I., and Levine, S. Rvs: What is essential for offline RL via supervised learning? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=S874XAIpkR-.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., Legg, S., and Kavukcuoglu, K. IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1407–1416. PMLR, July 2018. URL https://proceedings.mlr.press/v80/espeholt18a.html.

Fang, H., Duan, J., Clay, D., Wang, S., Liu, S., Huang, W., Fan, X., Tsai, W.-C., Chen, S., Wang, Y. R., Xing, S., Cho, J., Park, J. S., Eftekhar, A., Sushko, P., Farley, K., Wadhwa, A., Harrison, C., Han, W., Lee, Y.-C., VanderBilt, E., Hendrix, R., Ellawela, S., Ngoo, L., Chai, J., Ren, Z., Farhadi, A., Fox, D., and Krishna, R. MolmoAct2: Action Reasoning Models for Real-world Deployment, May 2026. URL http://arxiv.org/abs/2605.02881. arXiv:2605.02881 [cs] version: 1.

Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. D4RL: Datasets for Deep Data-Driven Reinforcement Learning, February 2021. URL http://arxiv.org/abs/2004.07219. arXiv:2004.07219 [cs].

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1861–1870. PMLR, July 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Hu, Z., Wu, R., Enock, N., Li, J., Kadakia, R., Erickson, Z., and Kumar, A. RaC: Robot Learning for Long-Horizon Tasks by Scaling Recovery and Correction, September 2025. URL http://arxiv.org/abs/2509.07953. arXiv:2509.07953 [cs].

Intelligence, P., Ai, B., Amin, A., Aniceto, R., Balakrishna, A., Balke, G., Black, K., Bokinsky, G., Cao, S., Charbonnier, T., Choudhary, V., Collins, F., Conley, K., Connors, G., Darpinian, J., Dhabalia, K., Dhaka, M., DiCarlo, J., Driess, D., Equi, M., Esmail, A., Fang, Y., Finn, C., Glossop, C., Godden, T., Goryachev, I., Groom, L., Habeeb, H., Hancock, H., Hausman, K., Hussein, G., Hwang, V., Ichter, B., Jacobsen, C., Jakubczak, S., Jen, R., Jones, T., Kammerer, G., Katz, B., Ke, L., Khadikov, M., Kuchi, C., Lamb, M., LeBlanc, D., LeCount, B., Levine, S., Li, X., Li-Bell, A., Lialin, V., Liang, Z., Lim, W., Lu, Y., Luo, E., Mano, V., Marwaha, N., Mongush, A., Murphy, L., Nair, S., Patterson, T., Pertsch, K., Ren, A. Z., Schelske, G., Sharma, C., Shi, B., Shi, L. X., Smith, L., Springenberg, J. T., Stachowicz, K., Stoeckle, W., Tang, J., Tanner, J., Tekeste, S., Torne, M., Vedder, K., Vuong, Q., Walling, A., Wang, H., Wang, J., Wang, X., Whalen, C., Whitmore, S., Williams, B., Xu, C., Yoo, S., Yu, L., Zhang, W., Zhang, Z., and Zhilinsky, U. $\pi_{0.7}$: a Steerable Generalist Robotic Foundation Model with Emergent Capabilities, April 2026. URL http://arxiv.org/abs/2604.15483. arXiv:2604.15483 [cs].

Kostrikov, I., Nair, A., and Levine, S. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=68n2s9ZJWF8.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative Q-Learning for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0d2b2061826a5df3221116a5085a6052-Abstract.html.

Lange, S., Gabel, T., and Riedmiller, M. Batch Reinforcement Learning. In Wiering, M. and van Otterlo, M. (eds.), *Reinforcement Learning: State-of-the-Art*, pp. 45–73. Springer, Berlin, Heidelberg, 2012. ISBN 978-3-642-27645-3. doi: 10.1007/978-3-642-27645-3_2. URL https://doi.org/10.1007/978-3-642-27645-3_2.

Le Roux, N., Bellemare, M., Lebensold, J., Bergeron, A., Greaves, J., Fréchette, A., Pelletier, C., Thibodeau-Laufer, E., Tóth, S., and Work, S. Tapered Off-Policy REINFORCE - Stable and efficient reinforcement learning for large language models. *Advances in Neural Information Processing Systems*, 38:68152–68180, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/6274d57365d7a6be06e58cad30d1b9da-Abstract-Conference.html.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems, November 2020. URL http://arxiv.org/abs/2005.01643. arXiv:2005.01643 [cs].

Li, H., Zuo, Y., Yu, J., Zhang, Y., Zhaohui, Y., Zhang, K., Zhu, X., Zhang, Y., Chen, T., Cui, G., Wang, D., Luo, D., Fan, Y., Sun, Y., Zeng, J., Pang, J., Zhang, S., Wang, Y., Mu, Y., Zhou, B., and Ding, N. SimpleVLA-RL: Scaling VLA training via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=TQhSodCM4r.

Li, Q., Kumar, A., Kostrikov, I., and Levine, S. Efficient deep reinforcement learning requires regulating overfitting. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=14-kr46GvP-.

Liu, B., Liu, X., Jin, X., Stone, P., and Liu, Q. Conflict-Averse Gradient Descent for Multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18878–18890. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9d27fdf2477ffbff837d73ef7ae23db9-Abstract.html.

Mandlekar, A., Xu, D., Wong, J., Nasiriany, S., Wang, C., Kulkarni, R., Fei-Fei, L., Savarese, S., Zhu, Y., and Martín-Martín, R. What matters in learning from offline human demonstrations for robot manipulation. In *5th Annual Conference on Robot Learning*, 2021. URL https://openreview.net/forum?id=JrsfBJtDFdI.

McCloskey, M. and Cohen, N. J. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. 24:109–165, 1989. doi: 10.1016/S0079-7421(08)60536-8. URL https://linkinghub.elsevier.com/retrieve/pii/S0079742108605368. Book Title: Psychology of Learning and Motivation ISBN: 9780125433242.

MiniMax, Chen, A., Li, A., Gong, B., Jiang, B., Fei, B., Yang, B., Shan, B., Yu, C., Wang, C., Zhu, C., Xiao, C., Du, C., Zhang, C., Qiao, C., Zhang, C., Du, C., Guo, C., Chen, D., Ding, D., Sun, D., Li, D., Jiao, E., Zhou, H., Zhang, H., Ding, H., Sun, H., Feng, H., Cai, H., Zhu, H., Sun, J., Zhuang, J., Cai, J., Song, J., Zhu, J., Li, J., Tian, J., Liu, J., Xu, J., Yan, J., Liu, J., He, J., Feng, K., Yang, K., Xiao, K., Han, L., Wang, L., Yu, L., Feng, L., Li, L., Zheng, L., Du, L., Yang, L., Zeng, L., Yu, M., Tao, M., Chi, M., Zhang, M., Lin, M., Hu, N., Di, N., Gao, P., Li, P., Zhao, P., Ren, Q., Xu, Q., Li, Q., Wang, Q., Tian, R., Leng, R., Chen, S., Chen, S., Shi, S., Weng, S., Guan, S., Yu, S., Li, S., Zhu, S., Li, T., Cai, T., Liang, T., Cheng, W., Kong, W., Li, W., Chen, X., Song, X., Luo, X., Su, X., Li, X., Han, X., Hou, X., Lu, X., Zou, X., Shen, X., Gong, Y., Ma, Y., Wang, Y., Shi, Y., Zhong, Y., Duan, Y., Fu, Y., Hu, Y., Gao, Y., Fan, Y., Yang, Y., Li, Y., Hu, Y., Huang, Y., Li, Y., Xu, Y., Mao, Y., Shi, Y., Wenren, Y., Li, Z., Li, Z., Tian, Z., Zhu, Z., Fan, Z., Wu, Z., Xu, Z., Yu, Z., Lyu, Z., Jiang, Z., Gao, Z., Wu, Z., Song, Z., and Sun, Z. MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention, June 2025. URL https://arxiv.org/abs/2506.13585v1.

Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M. Safe and Efficient Off-Policy Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://papers.nips.cc/paper_files/paper/2016/hash/c3992e9a68c5ae12bd18488bc579b30d-Abstract.html.

Nauman, M., Bortkiewicz, M., Miłoś, P., Trzcinski, T., Ostaszewski, M., and Cygan, M. Overestimation, Overfitting, and Plasticity in Actor-Critic: the Bitter Lesson of Reinforcement Learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 37342–37364. PMLR, July 2024. URL https://proceedings.mlr.press/v235/nauman24a.html.

Park, S., Frans, K., Mann, D., Eysenbach, B., Kumar, A., and Levine, S. Horizon Reduction Makes RL Scalable, June 2025. URL http://arxiv.org/abs/2506.04168. arXiv:2506.04168 [cs].

Pomerleau, D. A. ALVINN: an autonomous land vehicle in a neural network. In *Proceedings of the 2nd International Conference on Neural Information Processing Systems*, NIPS'88, pp. 305–313, Cambridge, MA, USA, January 1988. MIT Press.

Precup, D., Sutton, R. S., and Singh, S. Eligibility Traces for Off-Policy Policy Evaluation. January 2000. URL https://hdl.handle.net/20.500.14394/10401.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Team, G. R., Abdolmaleki, A., Abeyruwan, S., Ainslie, J., Alayrac, J.-B., Arenas, M. G., Balakrishna, A., Batchelor, N., Bewley, A., Bingham, J., Bloesch, M., Bousmalis, K., Brakel, P., Brohan, A., Buschmann, T., Byravan, A., Cabi, S., Caluwaerts, K., Casarini, F., Chan, C., Chang, O., Chappellet-Volpini, L., Chen, J. E., Chen, X., Chiang, H.-T. L., Choromanski, K., Collister, A., D'Ambrosio, D. B., Dasari, S., Davchev, T., Dave,

M. K., Devin, C., Palo, N. D., Ding, T., Doersch, C., Dostmohamed, A., Du, Y., Dwibedi, D., Egambaram, S. T., Elabd, M., Erez, T., Fang, X., Fantacci, C., Fong, C., Frey, E., Fu, C., Gao, R., Giustina, M., Gopalakrishnan, K., Graesser, L., Groth, O., Gupta, A., Hafner, R., Hansen, S., Hasenclever, L., Haves, S., Heess, N., Hernaez, B., Hofer, A., Hsu, J., Huang, L., Huang, S. H., Iscen, A., Jacob, M. G., Jain, D., Jesmonth, S., Jindal, A., Julian, R., Kalashnikov, D., Karagozler, M. E., Karp, S., Kecman, M., Kew, J. C., Kim, D., Kim, F., Kim, J., Kipf, T., Kirmani, S., Konyushkova, K., Ku, L. Y., Kuang, Y., Lampe, T., Laurens, A., Le, T. A., Leal, I., Lee, A. X., Lee, T.-W. E., Lever, G., Liang, J., Lin, L.-H., Liu, F., Long, S., Lu, C., Maddineni, S., Majumdar, A., Maninis, K.-K., Marmon, A., Martinez, S., Michaely, A. H., Milonopoulos, N., Moore, J., Moreno, R., Neunert, M., Nori, F., Ortiz, J., Oslund, K., Parada, C., Parisotto, E., Paryag, A., Pooley, A., Power, T., Quaglino, A., Qureshi, H., Raju, R. V., Ran, H., Rao, D., Rao, K., Reid, I., Rendleman, D., Reymann, K., Rivas, M., Romano, F., Rubanova, Y., Sampedro, P. P., Sanketi, P. R., Shah, D., Sharma, M., Shea, K., Shridhar, M., Shu, C., Sindhwani, V., Singh, S., Soricut, R., Sterneck, R., Storz, I., Surdulescu, R., Tan, J., Tompson, J., Tunyasuvunakool, S., Varley, J., Vesom, G., Vezzani, G., Villalonga, M. B., Vinyals, O., Wagner, R., Wahid, A., Welker, S., Wohlhart, P., Wu, C., Wulfmeier, M., Xia, F., Xiao, T., Xie, A., Xie, J., Xu, P., Xu, S., Xu, Y., Xu, Z., Yan, J., Yang, S., Yang, S., Yang, Y., Yu, H. H., Yu, W., Yuan, W., Yuan, Y., Zhang, J., Zhang, T., Zhang, Z., Zhou, A., Zhou, G., and Zhou, Y. Gemini Robotics 1.5: Pushing the Frontier of Generalist Robots with Advanced Embodied Reasoning, Thinking, and Motion Transfer, November 2025. URL http://arxiv.org/abs/2510.03342. arXiv:2510.03342 [cs].

Wang, S., Yu, L., Gao, C., Zheng, C., Liu, S., Lu, R., Dang, K., Chen, X.-H., Yang, J., Zhang, Z., Liu, Y., Yang, A., Zhao, A., Yue, Y., Song, S., Yu, B., Huang, G., and Lin, J. Beyond the 80/20 Rule: High-Entropy Minority Tokens Drive Effective Reinforcement Learning for LLM Reasoning. *Advances in Neural Information Processing Systems*, 38: 115452–115486, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/a797c2d2e0c1fdabf4d1ab8cd0b465c6-Abstract-Conference.html.

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., and Finn, C. Gradient surgery for multi-task learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in neural information processing systems*, volume 33, pp. 5824–5836. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/3fe78a8acf5fda99de95303940a2420c-Paper.pdf.