

Return-Weighted ELBO Fine-Tuning Degrades Masked Diffusion Planners



Presented by

Muhammad Ali Khan

muhammad.khan.25@ucl.ac.uk

Muhammad Ali Khan*, Mathis Weil*, Ahmet H. Güzel, Jack Parker-Holder, Ilija Bogunovic

University College London University of Basel *Equal contribution

NeurIPS 2026 Workshop: Beyond Next-Token Prediction

Policy gradients need a likelihood that masked diffusion lacks

- Masked discrete diffusion models generate a sequence by iteratively denoising it, so a plan can be revised globally rather than committed to one token at a time.
- Policy gradients require $\nabla_\theta \log \pi_\theta(\tau)$, and a masked diffusion model has no closed-form sequence log-likelihood.
- The tractable substitute is the return-weighted ELBO, which scales each sampled trajectory's denoising loss by a function of its return.
- It is a small change to a behavioural cloning trainer.

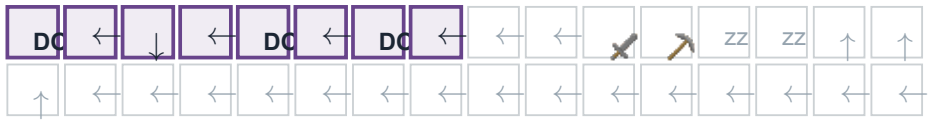
In our experiments it also degrades the model it starts from.

We test whether the return signal is responsible, and the evidence does not support it.

A ReMDM planner, pretrained by imitation, then fine-tuned

- A ReMDM planner [Wang et al., 2026b] denoises a 32-action plan on Craftax Classic [Matthews et al., 2024], executes a prefix and replans.
- Pretrained with DAgger [Ross et al., 2011], against a PPO-RNN teacher in Craftax Classic and a BFS oracle in MiniHack [Samvelyan et al., 2021].
- 25 conditions, three seeds each, 500 iterations, all from the same checkpoint (9.33M parameters on Craftax Classic).
- With three seeds the smallest attainable p is 0.10, so we report ranges, rank correlations and bootstrap intervals.





The checkpoint's plan from this frame. At evaluation the first 8 actions (outlined) are executed before replanning.

The objective weights each plan window by its own return

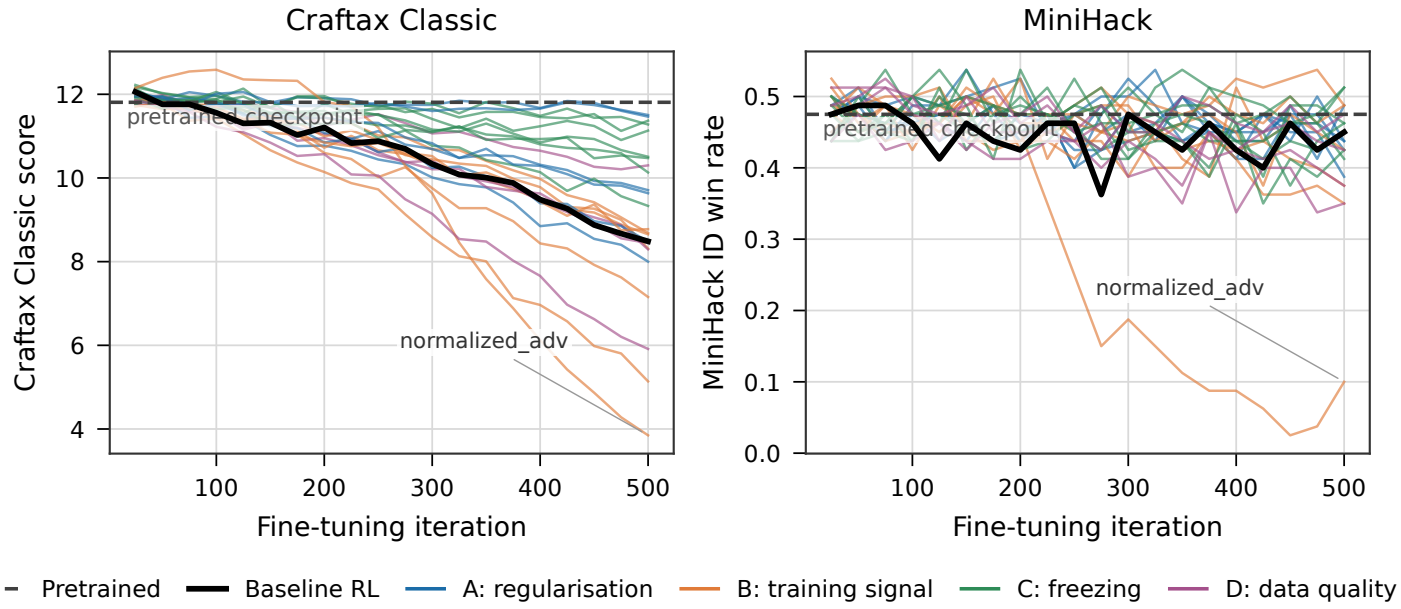
Each window's return R_i sums the reward over exactly the H steps it trains on.

$$A_i \;=\; \mathrm{clip}\left(\frac{\max(R_i, 0)}{\mu_{\text{batch}} + \epsilon},\; 0.1,\; 5.0\right), \qquad \mathcal{L}_{\text{RW}}(\theta) \;=\; \frac{1}{B}\sum_i A_i\, \ell_\theta(\boldsymbol{a}_i, o_i)$$

- ℓ_θ is the per-window ELBO, a masked cross-entropy weighted by the noise schedule.
- It is the positive-advantage branch of SPG [Wang et al., 2026a], with clipped positive returns in place of a group-relative advantage.
- Every weight is non-negative, so the objective can re-rank what the policy already produces but never push mass away from it.

Fine-tuning degrades the checkpoint

- On Craftax Classic the checkpoint scores 11.81. After 500 iterations it scores 8.22 (±0.14).
- None of the 25 conditions ends above the checkpoint. The best is LoRA at 11.63 (±0.03).
- MiniHack moves less, 47.5% to 43.8% (±6.1), and we do not read its three nominal gains as improvements.



All 25 conditions, coloured by group, with the pretrained checkpoint dashed.

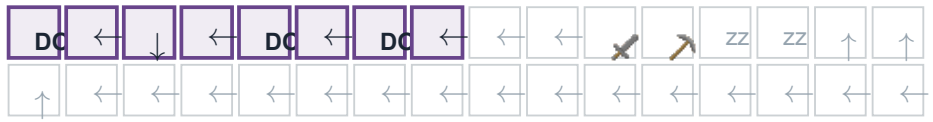
In one world, the fine-tuned plan repeats a single action

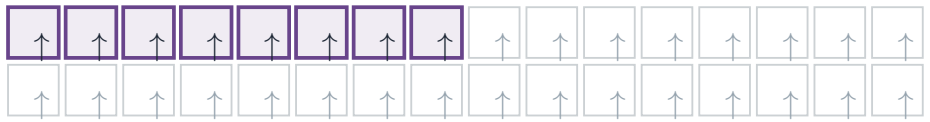
Checkpoint



After 500 iterations



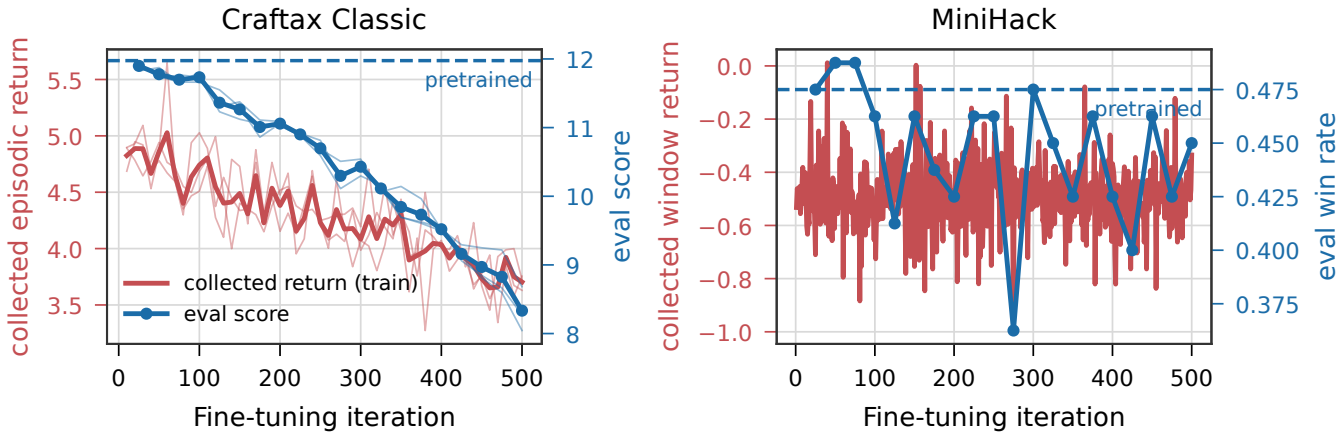




One Craftax Classic world and seed at step 64, rendered by the environment. Each planner denoises a 32-action plan from the frame shown. A single example, shown for illustration.

Collected return falls alongside the held-out score

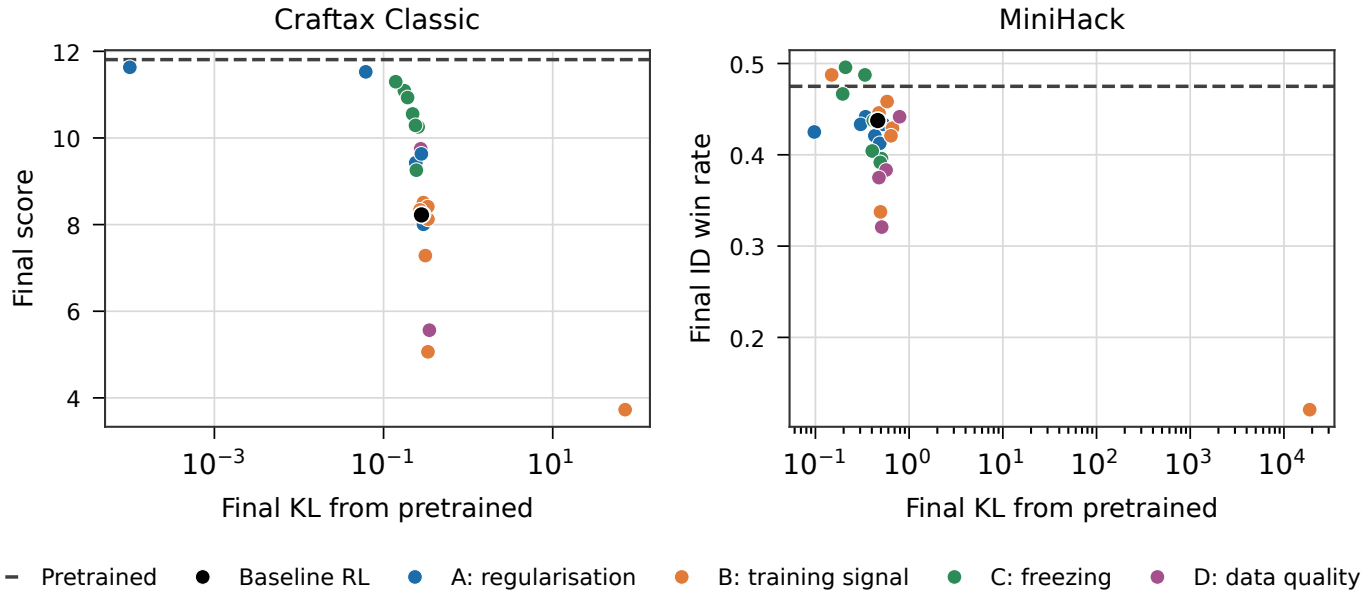
- The weights are computed from collection-time return, so a successful re-ranking should raise it.
- On Craftax Classic it falls from 4.83 to 3.70 while the held-out score falls from 11.90 to 8.33.
- The objective is not raising its own return while failing the held-out one.



Collection-time return (red, left axis) against held-out score (blue, right axis) for baseline RL, with the checkpoint dashed. On MiniHack both are flat.

Outcomes track how far each condition may move

- The top five on Craftax Classic are the five that most restrict the update, LoRA, trust-region KL, head-only, frozen backbone and top-1 layer.
- Final KL from the checkpoint correlates with final score at Spearman −0.92 on Craftax Classic and −0.48 on MiniHack.
- Distance alone does not account for it. A random displacement to the same KL scores 3.14 against the objective's 8.43.



Final score against final KL from the checkpoint, log scale, with the checkpoint dashed.

An exact decomposition isolates the return term

With $\delta_i = A_i/\bar{A} - 1$, so that $\sum_i \delta_i = 0$, the gradient factors with no approximation.

$$\nabla_\theta \mathcal{L}_{\mathrm{RW}} \;=\; \bar{A}\left(\underbrace{\nabla_\theta \mathcal{L}_{\mathrm{BC}}}_{\text{imitation}} \;+\; \underbrace{\tfrac{1}{B}\textstyle\sum_i \delta_i \nabla_\theta \ell_\theta(\boldsymbol{a}_i, o_i)}_{g_\delta \text{ (return term)}}\right)$$

- $\bar{A}$ rescales the step and leaves the direction unchanged, so everything the return does enters through g_δ.
- At the checkpoint $\|g_\delta\|/\|\nabla \mathcal{L}_{\mathrm{BC}}\| = 0.48 \pm 0.01$, about half the imitation gradient in norm, measured on one batch with two backward passes.
- A shuffled-δ null reproduces 0.74 of the ratio, so its size mostly reflects weight dispersion.

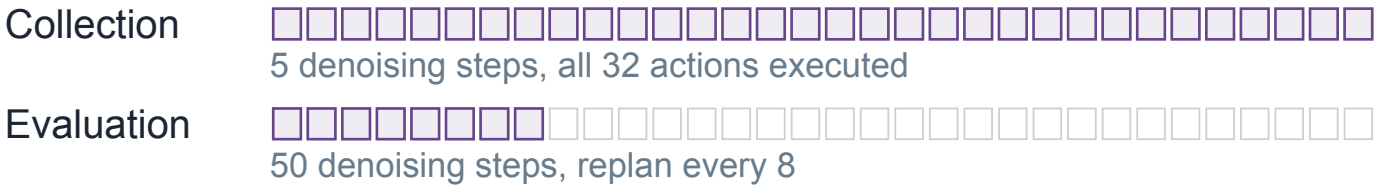
Shrinking the return term does not help

- Advantage clipping cuts the return term fivefold and scores 3.16 below baseline RL, second worst in the suite.
- Across the three weight transforms, the larger the return term, the higher the score. A return-driven degradation predicts the reverse.
- Clipping also ends further from the checkpoint, so displacement is a live confound.
- Removing the weighting costs more. With collection frozen at the checkpoint, uniform weights lose 2.45 points more.

Weight transform	$\|g_\delta\|/\|\nabla\mathcal{L}_{\mathrm{BC}}\|$	KL drift	Score
Clipped return ratio	0.48 ± 0.01	0.281	8.22
BC on wins	0.417 ± 0.025	0.312	7.28
Advantage clip	0.097 ± 0.003	0.335	5.06

Return term at the checkpoint on Craftax Classic. Advantage clip confines the weights to [0.8, 1.2]. Score and KL drift are three-seed means.

Matching the sampler removes most of the degradation



- Collected at the evaluation settings, baseline RL drops 0.26 instead of 3.68.
- A continued-DAgger arm relabels each visited state with the teacher and loses 1.70 and 0.31, so the checkpoint is not undertrained.
- The settings differ in denoising steps and in replanning, which we do not separate.

Change in score	Original	Evaluation
Return-weighted ELBO	-3.68 ± 0.19	-0.26 ± 0.16
Continued DAgger	-1.70 ± 0.14	-0.31 ± 0.09

Craftax Classic, three seeds, 500 iterations. Each drop is against the same run's evaluation of the checkpoint.

What we conclude

- Return-weighted ELBO fine-tuning degrades an imitation-pretrained masked diffusion planner, and we are unable to attribute that degradation to the return weighting.
- Most of it lies in the sampling mismatch rather than the return weighting, on one benchmark and one model.
- The weighting looks protective rather than causal. What it protects against is open.
- We recommend reporting CV_A and $\|g_\delta\|/\|\nabla \mathcal{L}_{\mathrm{BC}}\|$ before applying this objective. Each costs at most one batch and no accelerator.

Paper on OpenReview

References

Michael Matthews, Michael Beukman, Benjamin Ellis, Mikayel Samvelyan, Matthew Jackson, Samuel Coward, and Jakob Foerster. Craftax: A Lightning-Fast Benchmark for Open-Ended Reinforcement Learning, June 2024.

Stephane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning, March 2011.

Mikayel Samvelyan, Robert Kirk, Vitaly Kurin, Jack Parker-Holder, Minqi Jiang, Eric Hambro, Fabio Petroni, Heinrich Küttler, Edward Grefenstette, and Tim Rocktäschel. MiniHack the Planet: A Sandbox for Open-Ended Reinforcement Learning Research, November 2021.

Chenyu Wang, Paria Rashidinejad, DiJia Su, Song Jiang, Sid Wang, Siyan Zhao, Cai Zhou, Shannon Zejiang Shen, Feiyu Chen, Tommi Jaakkola, Yuandong Tian, and Bo Liu. SPG: Sandwiched Policy Gradient for Masked Diffusion Language Models, April 2026a.

Guanghan Wang, Yair Schiff, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Remasking Discrete Diffusion Models with Inference-Time Scaling, February 2026b.